

Nonparametric identification and maximum likelihood estimation for hidden Markov models

Grigory Alexandrovich, Hajo Holzmann¹ and Anna Leister

Fakultät für Mathematik und Informatik, Philipps-Universität Marburg, Germany

Nonparametric identification and maximum likelihood estimation for finite-state hidden Markov models are investigated. We obtain identification of the parameters as well as the order of the Markov chain if the transition probability matrices have full-rank and are ergodic, and if the state-dependent distributions are all distinct, but not necessarily linearly independent. Based on this identification result, we develop nonparametric maximum likelihood estimation theory. First, we show that the asymptotic contrast, the Kullback–Leibler divergence of the hidden Markov model, identifies the true parameter vector nonparametrically as well. Second, for classes of state-dependent densities which are arbitrary mixtures of a parametric family, we show consistency of the nonparametric maximum likelihood estimator. Here, identification of the mixing distributions need not be assumed. Numerical properties of the estimates as well as of nonparametric goodness of fit tests are investigated in a simulation study.

Keywords: hidden Markov models, latent state models, nonparametric identification, nonparametric maximum likelihood estimation

1. Introduction

A discrete-time hidden Markov model consists of an observed process $(Y_t)_{t \in \mathbb{N}}$ as well as a latent, unobserved process $(X_t)_{t \in \mathbb{N}}$, such that the Y_t are independent given the X_t , the conditional distribution of Y_s given the X_t depends on X_s only and X_t is a finite-state Markov chain. We assume that X_t is time-homogeneous. The cardinality K of the state space of X_t is called the number of states. The conditional distributions of Y_s given $X_s = k$ ($k = 1, \dots, K$) are called the state-dependent distributions, and we assume that they are independent of s . The entries of the transition probability matrix are denoted by $\Gamma = (\alpha_{j,k})_{j,k=1,\dots,K}$. Further, we assume that the Y_t take values in any subset of Euclidean space $\mathcal{S} \subset \mathbb{R}^q$, and denote the distribution functions of the state-dependent distributions by F_k ($k = 1, \dots, K$).

Parametric estimation theory for finite-state hidden Markov models is well-developed; see Leroux [14] for consistency and Bickel et al. [2] for asymptotic normality of the maximum likelihood estimator. In order to achieve greater flexibility and to avoid misspecification,

¹Address for correspondence: Prof. Dr. Hajo Holzmann, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerweinstr. D-35032 Marburg, Germany email: holzmann@mathematik.uni-marburg.de, Fon: + 49 6421 2825454

nonparametric modelling and estimation of the component distributions have received recent interest, see Dannemann et al. [3], Gassiat et al. [5] and Vernet [20]. However, the most basic question is whether such models are identifiable. We give an affirmative answer in great generality: if the transition probability matrix Γ is ergodic and of full rank, and if the state-dependent distributions are all distinct, then the parameters, together with the number of states, are all identifiable. Our second main result states that the asymptotic contrast for maximum likelihood estimation, the generalized Kullback–Leibler divergence of the hidden Markov model, uniquely identifies the true parameter nonparametrically. It is well known that the ordinary Kullback–Leibler divergence discriminates between any two probability distributions on the same measurable space without reference to a particular model. However, for hidden Markov models, for which the generalized Kullback–Leibler divergence is defined as a limit of normalized log-likelihoods, the contrast property had previously been deduced from mere parametric identification of mixtures of product distributions in Leroux [14], and had thus been restricted to parametric settings. Our second result allows us to investigate consistency of the maximum likelihood estimator over nonparametric classes. As an important example, we consider general mixtures of a parametric family as model for the state-dependent distributions, and as a third main result obtain consistency of the mixture densities under suitable assumptions. Here, we do not assume that the mixing distributions themselves are identified, and thus allow, e.g., general mixtures of normals.

Let us discuss how our identification results relate to previous ones in the literature. In a seminal paper, based on a result by Kruskal [12] on the identification of factors in three-way tables, Allman et al. [1] showed generic identifiability of various latent-state models, including hidden Markov models with finite-valued observations. Strict point identification, up to label swapping, for general-valued hidden Markov models was recently discussed by Gassiat et al. [5] and Gassiat and Rousseau [6]. Using analytic arguments, Gassiat and Rousseau [6] showed that if Γ has full rank, and if the state-dependent distributions are from a location family of an arbitrary density, then all parameters as well as the number of components are identified from the joint distribution of two observations. While certainly of interest, merely the assumption of equal scale in each component which is implied by the model may be too restrictive for most applications. For a given K , Gassiat et al. [5] show identification if Γ has full-rank and if the state-dependent distributions are linearly independent. The result follows by combining arguments given in Allman et al. [1] for generic identification of hidden Markov models and of finite mixtures of product distributions. While the assumption of linearly independent state-dependent distributions is convenient in the proofs, it is not intuitive, and also difficult to interpret for nonparametric classes such as smooth classes of densities, or shape-constrained classes such as log-concave densities, where more than two distinct distributions may well be linearly dependent. Our result for distinct state-dependent distributions is better suited for such nonparametric classes. In its proof, the main challenge is to find a substitute for the linear independence of the state-dependent distributions.

2. Nonparametric identification

Our basic assumptions are as follows.

A1. The transition probability matrix $\Gamma = (\alpha_{j,k})_{j,k=1,\dots,K}$ of (X_t) has full rank and is ergodic.

A2. The state-dependent distributions F_k ($k = 1, \dots, K$) are all distinct.

Let us first consider the stationary case for a fixed number of components.

A3. The Markov chain (X_t) is stationary with starting distribution π , the stationary distribution of Γ .

Theorem 1. *For given K , let Γ, F_1, \dots, F_K and $\tilde{\Gamma}, \tilde{F}_1, \dots, \tilde{F}_K$ be two sets of parameters for a K -state hidden Markov model, such that the joint distributions of (Y_1, \dots, Y_{2K+1}) under both sets of parameters are equal. Further, suppose that Γ and F_1, \dots, F_K satisfy Assumptions A1–A3. Then both sets of parameters coincide up to label swapping.*

In Theorem 1, Assumptions A1 and A2 solely concern Γ, F_1, \dots, F_K ; nothing is assumed for $\tilde{\Gamma}, \tilde{F}_1, \dots, \tilde{F}_K$. For statistical inference, this implies that Assumptions A1–A3 are required for the true model, but estimators need not be restricted to satisfy these constraints.

Example 1. To show the necessity of the full-rank assumption of the transition probability matrix, we construct for each $K \geq 1$ a $(K+1)$ -state matrix of rank K and two sets of $K+1$ distributions, which are even linearly independent, such that the observations in a resulting $(K+1)$ -state hidden Markov model have the same distribution. To this end, let $\Gamma = (\alpha_{j,k})_{j,k=1,\dots,K}$ be a K -state ergodic transition probability matrix of full rank. Let $\delta, \beta \in (0, 1)$ with $\delta \neq \beta$, set $p = \beta/(1 + \beta - \delta)$ for which $p \in (0, 1)$, and consider the $(K+1)$ -state matrix of rank K

$$\Gamma_1 = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,K-1} & p\alpha_{1,K} & (1-p)\alpha_{1,K} \\ \vdots & & \vdots & & \vdots \\ \alpha_{K-1,1} & \cdots & \alpha_{K-1,K-1} & p\alpha_{K-1,K} & (1-p)\alpha_{K-1,K} \\ \alpha_{K,1} & \cdots & \alpha_{K,K-1} & p\alpha_{K,K} & (1-p)\alpha_{K,K} \\ \alpha_{K,1} & \cdots & \alpha_{K,K-1} & p\alpha_{K,K} & (1-p)\alpha_{K,K} \end{pmatrix}.$$

Let F_1, \dots, F_{K+1} be linearly independent distribution functions, for example, normal distributions with distinct parameters. As the second set $\tilde{F}_1, \dots, \tilde{F}_{K+1}$ of distribution functions let $\tilde{F}_1 = F_1, \dots, \tilde{F}_{K-1} = F_{K-1}$ and

$$\tilde{F}_K = \delta F_K + (1 - \delta) F_{K+1}, \quad \tilde{F}_{K+1} = \beta F_K + (1 - \beta) F_{K+1}.$$

Then $p\tilde{F}_K + (1 - p)\tilde{F}_{K+1} = pF_K + (1 - p)F_{K+1}$, and from Holzmam and Schwaiger [9], the distributions of the observations of a $(K+1)$ -state hidden Markov model with transition probability matrix Γ_1 , stationary starting distribution and either set of state-dependent distributions are equal to that of a K -state stationary hidden Markov model with transition probability matrix Γ and state dependent distributions F_1, \dots, F_{K-1} and $pF_K + (1 - p)F_{K+1}$.

Example 2. One may wonder whether the assumption of distinct state-dependent distributions is actually necessary for identification, or whether states may possibly be reconstructed merely from transitions if there are sufficiently many different state-dependent

distributions, without all of them being distinct. In this example we describe a class of hidden Markov models where this is not possible. A stationary Markov chain $(X_t)_{t \in \mathbb{N}}$ with transition probability matrix Γ is called lumpable with respect to a partition $\{G_1, \dots, G_m\}$ of the state-space $\{1, \dots, K\}$ if the process $(\tilde{X}_t)_{t \in \mathbb{N}}$ defined by $\tilde{X}_t = j$ if $X_t \in G_j$ ($j = 1, \dots, m$) is also a Markov chain. Kemény and Snell [10] show that this is equivalent to $\text{pr}(X_{t+1} \in G_j \mid X_t \in G_i) = \text{pr}(X_{t+1} \in G_j \mid X_t = k)$ ($i, j = 1, \dots, m; k \in G_i$). If this is the case in a hidden Markov model $(Y_t, X_t)_{t \in \mathbb{N}}$, for which the state-dependent distributions are equal over the states in the elements G_j ($j = 1, \dots, m$) of the partition, then its distribution reduces to that of a m -state hidden Markov model. In particular, the full transition probability matrix Γ of the K -state representation cannot be identified.

Example 3. Hidden Markov models with state-dependent densities which mainly differ in terms of their scale are used for modelling financial time series. The distinct scales correspond to volatility states of the market; see Holmann and Schwaiger [9] and references therein. In case of three states, there is a transition regime between highest and lowest volatility. Hence it is plausible that the state-dependent density with intermediate scale may actually be a mixture of the two densities with lowest and highest volatility, thus making the three state-dependent densities linearly dependent. See the supplementary material for simulations in such a scenario.

Now let us turn to the case of a general starting distribution. While only of moderate statistical interest by itself, identification of the initial distribution is an essential tool for proving the nonparametric contrast property of the Kullback–Leibler divergence of a hidden Markov model in Section 3.1. The choice of T in the following theorem is due to the fact that for $t_0 = K^2 - 2K + 2$, Γ^{t_0} has strictly positive entries [8].

Theorem 2. *For a known number of states K , let $\lambda, \Gamma, F_1, \dots, F_K$ and $\tilde{\lambda}, \tilde{\Gamma}, \tilde{F}_1, \dots, \tilde{F}_K$ be two sets of parameters for a K -state hidden Markov model, where λ and $\tilde{\lambda}$ denote the initial distributions of the Markov chain. Suppose that the joint distributions of (Y_1, \dots, Y_T) with $T = (2K + 1)(K^2 - 2K + 2) + 1$, are equal under both sets of parameters. Further, suppose that Γ and F_1, \dots, F_K satisfy Assumptions A1 and A2. Then both sets of parameters coincide up to label swapping.*

Finally, let us turn to the additional identification of the number of states. For $L < K$ we may interpret an L -state as a K -state hidden Markov model, where $K - L$ states are never visited by the underlying Markov chain. From Theorem 2, we therefore get the following corollary.

Corollary 3. *Let λ, Γ and F_1, \dots, F_K and $\bar{\lambda}, \bar{\Gamma}$ and $\bar{F}_1, \dots, \bar{F}_L$ be two sets of parameters for a K -state and a L -state hidden Markov model, where $L \leq K$. Assume that Γ is ergodic and of full rank, and that F_1, \dots, F_K are all distinct. If the joint distributions of (Y_1, \dots, Y_T) , $T = (2K + 1)(K^2 - 2K + 2) + 1$, are the same under the both sets of parameters, then $K = L$ and the sets of parameters are equal up to a label swapping.*

In summary, we get the following identification result for the number of states and the parameters.

Corollary 4. *For a hidden Markov model, within the class of parameters satisfying Assumptions A1 and A2, both the number of states and the parameters are identified from the distribution of the observed process $(Y_t)_{t \in \mathbb{N}}$.*

Indeed, if we compare two hidden Markov models with L and K states satisfying Assumptions A1 and A2 and having equal distributions of the observations, then Theorem 2 takes care of the case $L = K$ while Corollary 3 shows that the case $L \neq K$ cannot occur.

3. Nonparametric maximum likelihood estimation

3.1. The Kullback–Leibler divergence of a hidden Markov model

Let \mathcal{D} be a class of densities on \mathcal{S} with respect to some σ -finite measure ν . Suppose that $(Y_t, X_t)_{t \in \mathbb{N}}$ is a K -state hidden Markov model with transition probability matrix Γ_0 satisfying Assumptions A1 and A3 and having stationary distribution π_0 , and that the state-dependent distributions $F_{1,0}, \dots, F_{K,0}$ are all distinct and have densities $f_{1,0}, \dots, f_{K,0}$ from the class \mathcal{D} . In the following, we write $Y_s^t = (Y_s, \dots, Y_t)$ and $y_s^t = (y_s, \dots, y_t)$ ($1 \leq s < t < \infty$).

For parameters $\lambda, \Gamma, f_1, \dots, f_K$, $n \in \mathbb{N}$ and $y_1^n \in \mathcal{S}^n$ consider

$$g_n(y_1^n; \lambda, \Gamma, f_1, \dots, f_K) = \sum_{x_1=1}^K \cdots \sum_{x_n=1}^K \lambda_{x_1} f_{x_1}(y_1) \prod_{i=2}^n \alpha_{x_{i-1}, x_i} f_{x_i}(y_i),$$

the joint density of n observations under these parameters, and denote the log-likelihood function of Y_1, \dots, Y_n by

$$L_n(\lambda, \Gamma, f_1, \dots, f_K) = \log g_n(Y_1^n; \lambda, \Gamma, f_1, \dots, f_K).$$

A4. The true densities $f_{j,0} \in \mathcal{D}$ satisfy $E\{|\log f_{j,0}(Y_1)|\} < \infty$, ($j = 1, \dots, K$).

A5. The model satisfies $E\{\log f(Y_1)\}^+ < \infty$, ($f \in \mathcal{D}$).

Theorem 5. Suppose that $(Y_t, X_t)_{t \in \mathbb{N}}$ is a K -state hidden Markov model with transition probability matrix Γ_0 satisfying Assumptions A1 and A3, and that the state-dependent distributions $F_{1,0}, \dots, F_{K,0}$ are all distinct and have densities $f_{1,0}, \dots, f_{K,0}$ from the class \mathcal{D} , and satisfy Assumption A4. Let λ, λ_0 be K -state probability vectors with strictly positive entries. Under Assumption A5, given $f_1, \dots, f_K \in \mathcal{D}$ we have almost surely as $n \rightarrow \infty$ that

$$\begin{aligned} n^{-1} \{L_n(\lambda, \Gamma, f_1, \dots, f_K) - L_n(\lambda_0, \Gamma_0, f_{1,0}, \dots, f_{K,0})\} \\ \rightarrow -K\{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} \in (-\infty, 0], \end{aligned} \quad (1)$$

and $K\{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} = 0$ if and only if the two sets of parameters are equal up to label swapping.

The limit in (1) defines the Kullback–Leibler divergence of the hidden Markov model. As could be expected, it does not identify the initial distribution, and arbitrary probability vectors with positive entries, not necessarily the stationary distribution of Γ , can be used in the likelihood function. It is well known that the ordinary Kullback–Leibler divergence discriminates between any two probability distributions on the same measurable space without reference to a particular model. For hidden Markov models, Leroux [14] showed that the limit in (1) may be represented as an integral over the ordinary

Kullback–Leibler divergence of finite segments of hidden Markov models, where integration is with respect to the initial distributions. From this and parametric identification of finite mixtures of product distributions, he deduced the contrast property, that is $K\{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} \in [0, \infty)$ and $K\{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} = 0$ if and only if the two sets of parameters are equal up to label swapping, within a parametric class. However, Theorem 2 implies that it holds without reference to a parametric family.

3.2. Nonparametric maximum likelihood estimation for state-dependent mixtures

In this subsection we use arbitrary mixtures of some parametric family of state-dependent densities to illustrate how the above results can be employed for nonparametric maximum likelihood estimation in hidden Markov models. Suppose that $(f_\vartheta)_{\vartheta \in \Theta}$ is a parametric family of densities on \mathcal{S} with respect to some σ -finite measure, and that $\Theta \subset \mathbb{R}^d$ is compact. Let $\tilde{\Theta}$ be the set of Borel probability measures on Θ . Endowed with the weak topology $\tilde{\Theta}$ is also a compact set. Assume that the map $(y, \vartheta) \mapsto f_\vartheta(y)$ is continuous on $\mathcal{S} \times \Theta$. Given $\mu \in \tilde{\Theta}$, we let

$$f_\mu(y) = \int_{\Theta} f_\vartheta(y) d\mu(\vartheta)$$

denote the corresponding mixture density. We shall call μ the mixing distribution for f_μ , and take $\mathcal{D} = \{f_\mu : \mu \in \tilde{\Theta}\}$ as model for the state-dependent densities.

For independent identically distributed observations there is some literature on nonparametric estimation of μ or f_μ . Lindsay [16] shows that there exists a nonparametric maximum likelihood estimator for μ with finite support, the number of support points being at most equal to the sample size. Leroux [15] obtains its consistency under the assumption that the mixing distribution μ is identified from f_μ . While convergence of estimators of μ may be quite slow [17], f_μ is estimated at optimal near-parametric rates for normal mixtures [11, 7]. We shall focus on consistency and in contrast to Leroux [15] do not assume that the mixing distribution μ is identified from the mixture density f_μ , since our interest is in the estimation of f_μ rather than of μ . Thus we allow, e.g., arbitrary mixtures of normal densities in both mean and variance, for which the mixing distribution is not identified [18].

Let $\theta = (\Gamma, \mu_1, \dots, \mu_K) \in G \times \tilde{\Theta} \times \dots \times \tilde{\Theta}$, where G is the compact set of K -state transition probability matrices. Given the sequence of observations Y_1, \dots, Y_n , the log-likelihood function is

$$L_n(\theta) = \log \left\{ \sum_{x_1=1}^K \cdots \sum_{x_n=1}^K \lambda_{x_1} f_{\mu_{x_1}}(Y_1) \prod_{i=2}^n \alpha_{x_{i-1}, x_i} f_{\mu_{x_i}}(Y_i) \right\},$$

where λ is an arbitrary K -state strictly positive probability vector. First, we show the existence of a maximum likelihood estimator for which the state-dependent mixing distributions μ_k ($k = 1, \dots, K$) have finite support.

Theorem 6. *Let $(f_\vartheta)_{\vartheta \in \Theta}$ be a parametric family of densities with $\Theta \subset \mathbb{R}^d$ compact, and let $\mathcal{D} = \{f_\mu : \mu \in \tilde{\Theta}\}$ be the model for the state-dependent densities, where $\tilde{\Theta}$ is the set of Borel probability measures on Θ . Then for any $n \geq 1$ there exists a maximum likelihood*

estimator $\hat{\theta}_n = (\hat{\Gamma}_n, \hat{\mu}_{1,n}, \dots, \hat{\mu}_{K,n})$, for which the state-dependent mixing distributions are of the form

$$\hat{\mu}_{k,n} = \sum_{j=1}^m a_j \delta_{\vartheta_{j,k}} \quad (k = 1, \dots, K),$$

where $m \in \{1, \dots, Kn + 1\}$, $a_j > 0$, $\sum_{j=1}^m a_j = 1$, $\vartheta_{j,k} \in \Theta$ ($j = 1, \dots, m$), and where δ_{ϑ} is the point-mass at ϑ .

This is similar to Lindsay [16]’s existence result, although more components are required due to the distinct states and the non-convexity of the likelihood of the hidden Markov model.

Let us turn to consistency. Assume that the true state-dependent densities $f_{k,0} = f_{\mu_{k,0}}$ belong to the model and are all distinct, and that Γ_0 satisfies Assumption A1.

A6. For every $\mu \in \tilde{\Theta}$ and a small enough neighborhood O_μ of μ we have

$$E \left[\sup_{\tilde{\mu} \in O_\mu} \{\log f_{\tilde{\mu}}(Y_1)\}^+ \right] < \infty.$$

Theorem 7. Let $(f_{\vartheta})_{\vartheta \in \Theta}$ be a parametric family of densities with $\Theta \subset \mathbb{R}^d$ compact, and let $\mathcal{D} = \{f_{\mu} : \mu \in \tilde{\Theta}\}$ be the model for the state-dependent densities, where $\tilde{\Theta}$ is the set of Borel probability measures on Θ . Suppose that Assumptions A1, A3, A4 and A6 hold, and let $\hat{\theta}_n = (\hat{\Gamma}_n, \hat{\mu}_{1,n}, \dots, \hat{\mu}_{K,n})$ denote a maximum likelihood estimator. Then, after relabeling, we have in probability as $n \rightarrow \infty$ that $\hat{\Gamma}_n \rightarrow \Gamma_0$ and

$$f_{\hat{\mu}_{k,n}}(y) \rightarrow f_{k,0}(y) \quad (y \in \mathcal{S}, k = 1, \dots, K).$$

Furthermore, if the mixing distribution μ is identified from the mixture density f_{μ} , then we additionally have that $d_w(\hat{\mu}_{k,n}, \mu_{k,0}) \rightarrow 0$ in probability, where d_w is a distance which metrizes weak convergence in $\tilde{\Theta}$.

4. Simulations

4.1. Nonparametric maximum likelihood estimation

In this section we investigate the performance of the nonparametric maximum likelihood estimator based on state-dependent mixtures in a simulation study, and discuss its applications to goodness of fit assessment of parametric models.

Consider a three-state hidden Markov model, in which the state-dependent densities are mixtures of univariate Gaussian distributions, specified as follows. Let $g_{\beta(a,b)}(x) = \{\Gamma(a+b)/\{\Gamma(a)\Gamma(b)\}x^{a-1}(1-x)^{b-1}\mathbb{1}_{(0,1)}(x)$ denote the density of the Beta distribution, $g_{\beta(a,b)}(x; l, s) = g_{\beta(a,b)}\{(x-l)/s\}/s$ the Beta density translated by l and scaled by s and let $\phi_{\mu,\sigma}$ denote the Gaussian density with parameters μ and σ . The state-dependent density of the first state is taken as $f_{1,0}(y) = 0.33\phi_{-10,2}(y) + 0.33\phi_{-7.5,2}(y) + 0.34\phi_{-4.2}(y)$. The densities of the second and third states, $f_{2,0}(y)$ and $f_{3,0}(y)$, are general mixtures of univariate Gaussian densities. For $f_{2,0}(y)$ we let μ follow the Beta distribution $g_{\beta(2,2)}(\mu; 0, 1)$ and let σ be uniformly distributed on the interval $(1, 4)$, for $f_{3,0}(y)$ we let μ follow $g_{\beta(2,11)}(\mu; 5, 33)$

and let σ be uniformly distributed on $(1.4, 1.6)$. For the transition probability matrix we choose

$$\Gamma = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.4 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}.$$

Computing the nonparametric maximum likelihood estimator of a mixing distribution and the resulting mixture is not an easy task, see, e.g., Laird [13]. We successively compute the maximum likelihood estimator for a given number of mixture components in each state using the expectation-maximization algorithm as described in Volant et al. [21], and increase the number of components as long as the resulting likelihood increases.

In our simulations, we use series of lengths $n = 1000$ from the above model. We also consider the maximum likelihood estimators in two misspecified parametric hidden Markov models with simple Gaussian and two-component mixtures of Gaussian distributions, respectively. The nonparametric maximum likelihood estimators are denoted by $f_{\hat{\mu}_{k,n}}$, the simple Gaussian estimators by $f_{\tilde{\mu}_{k,n}}$ and the two-component Gaussian mixture estimators by $f_{\bar{\mu}_{k,n}}$ ($k = 1, 2, 3$). On a computer with 3.07 GHz and 24GB RAM, computing the simple Gaussian estimators once requires 2.3 seconds, the two-component Gaussian mixture estimator requires 8.7 seconds and the nonparametric maximum likelihood estimators requires 84.5 seconds.

To illustrate the consistency of $f_{\hat{\mu}_{k,n}}$ as stated in Theorem 7, we evaluate the relative errors over 10000 simulations for the points indicated in Fig. 1 and listed in Table 1. The results together with those for the misspecified parametric estimators are given in Table 1. The relative errors for $f_{\tilde{\mu}_{k,n}}$ and $f_{\bar{\mu}_{k,n}}$ are higher at most points than those for $f_{\hat{\mu}_{k,n}}$, in particular for states 1 and 3, which reflects the bias of these estimators due to misspecification. The estimators for the transition probability matrices perform rather similarly for the three methods, therefore we do not report the results. Additional simulation results for series of lengths different from 1000, which illustrate the consistency of the nonparametric maximum likelihood estimator and its performance for shorter series, are provided in the supplementary material.

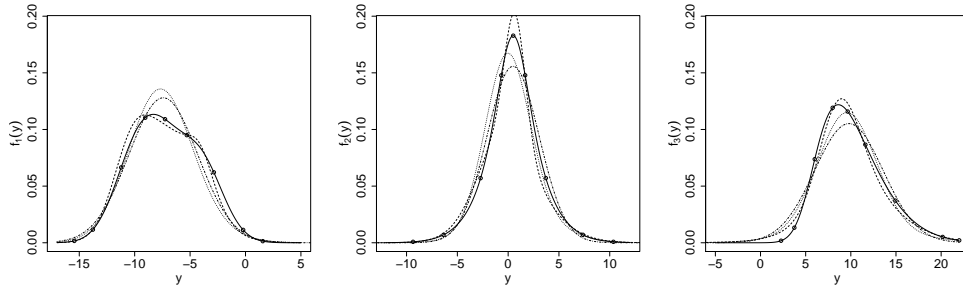


Figure 1: State-dependent densities and estimators for a typical sample. Solid line: true densities, dashed line: nonparametric maximum likelihood estimators, dotted line: two-component mixture maximum likelihood estimators, dot-dashed line: Gaussian maximum likelihood estimators

Figure 1 shows the state-dependent normal mixture densities $f_{k,0}$ as well as the fits $f_{\hat{\mu}_{k,n}}$, $f_{\tilde{\mu}_{k,n}}$ and $f_{\bar{\mu}_{k,n}}$ for a typical sample. The nonparametric estimator captures the overall

y	-15.45	-13.77	-11.22	-9.05	-7.26	-5.3	-2.86	-0.21	1.56
nonpar	109.79	28.00	6.92	12.94	23.93	5.09	43.82	46.40	43.87
2-comp	117.75	28.61	6.26	12.46	25.18	4.94	45.04	48.01	49.49
Gauss	136.66	31.14	5.84	10.68	24.37	4.68	43.15	52.93	37.43
y	-9.36	-6.36	-2.71	-0.68	0.5	1.67	3.71	7.36	10.36
nonpar	65.27	22.20	64.95	9.77	13.44	19.36	25.00	59.64	67.53
2-comp	69.44	22.63	68.76	10.60	13.88	19.48	25.12	61.55	67.06
Gauss	79.61	16.69	74.73	9.60	15.02	19.97	25.32	81.74	98.08
y	2.27	3.74	6	7.99	9.66	11.61	14.93	20.17	22
nonpar	1090.32	166.99	9.90	20.26	13.87	6.38	7.04	33.61	48.31
2-comp	1103.22	175.93	8.29	22.56	15.08	5.95	6.81	37.69	50.26
Gauss	1236.47	202.98	4.79	24.17	18.80	6.69	3.24	34.78	52.51

Table 1: Relative errors ($\times 100$) of the three estimators compared to the true densities at selected values for y averaged over 10000 replications. ‘Gauss’ stands for Gaussian state-dependent distributions, ‘2-comp’ for two component Gaussian mixtures and ‘nonpar’ for nonparametric Gaussian mixtures.

shape of the underlying density, in particular its skewness, much better than both parametric estimators, which deviate substantially from it.

4.2. Goodness of fit test

In this section we conduct a formal goodness of fit evaluation for a Gaussian hidden Markov model in the setting of Section 4.1. We use the likelihood ratio test against the nonparametric alternative of state-dependent general Gaussian mixtures as well as against the parametric alternative of state-dependent two-component Gaussian mixtures. Critical values are estimated by using the parametric bootstrap. Its consistency requires the asymptotic distribution to depend continuously on nuisance parameters, see van der Vaart [19], and caution is needed in irregular problems, see Drton and Williams [4].

To avoid excessive running times, we estimate critical values under the null model only once. First we use a single long series of the hidden Markov model with parameters as given in Section 4.1 and estimate the parameters under the null hypothesis of a Gaussian hidden Markov model, then from this estimated null model we simulate 10000 series of lengths 1000 to obtain the critical values against both classes of alternatives. Each simulated series requires about 85 seconds running time, so that on a computer with 12 processing units a total of 20 hours are required. Next, we also simulate from the model in Section 4.1 10000 series of lengths 1000 and use the simulated critical values to estimate the power.

The results for three significance levels are shown in Table 2. Although the critical values for the nonparametric test are larger, it still has slightly higher power at all three significance levels.

	Parametric vs two-component mixture	Parametric vs nonparametric
Critical value (90%)	1.72	2.42
Simulated power	95.67	96.44
Critical value (95%)	2.57	3.57
Simulated power	92.07	93.97
Critical value (99%)	4.33	6.03
Simulated power	79.83	85.27

Table 2: Simulated critical values and powers of the likelihood ratio tests

5. Discussion

We obtain nonparametric identification for hidden Markov models under assumptions that are close to minimal. In particular, linear independence of the state-dependent distributions is not required; they are merely assumed to be distinct. By example, we show the necessity of a full rank of the transition probability matrix. Ergodicity of Γ , which is assumed in most statistical estimation procedures in a parametric framework, is equivalent to irreducibility and aperiodicity. The proof of Theorem 1 does not require aperiodicity, so the conclusion holds without it, while the proof of Theorem 2 does require ergodicity. Irreducibility could potentially be dropped by considering communicating classes.

The majority of hidden Markov models used in applications have parametric state-dependent distributions. However, the need for distributions more flexible than the normal has been recognized in various papers. A popular alternative class are finite mixtures of normals with a given maximal number of components, see Volant et al. [21] and Holzmann and Schwaiger [9]. The fact that the nonparametric maximum likelihood estimator also has a finite number of components, although potentially growing with the sample size, makes the use of finite mixtures as state-dependent distributions even more attractive. Further, as demonstrated in the simulation section, comparison of parametric with nonparametric fits may be used for goodness of fit assessments, both formally by employing likelihood ratio tests and by visually comparing the parametric and nonparametric density estimates.

Acknowledgement

The authors would like to thank the editor, the associate editor and three reviewers for helpful comments which lead to improved contents and presentation. Anna Leister and Hajo Holzmann gratefully acknowledge financial support of the “Deutsche Forschungsgemeinschaft”.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the results and additional simulations for a scenario with linearly dependent state-dependent distributions.

A. Outline of the proofs

We present outlines of the proofs of the theorems.

Proof of Theorem 1. The proof follows that of Theorem 1 in Gassiat et al. [5], which in turn combines arguments in Allman et al. [1] for generic identification of hidden Markov models and finite mixtures of product distributions.

In the first step, for $T \geq K - 1$ we form the blocks

$$V_T = Y_1^T = (Y_1, \dots, Y_T), \quad W_T = Y_{T+2}^{2T+1} = (Y_{T+2}, \dots, Y_{2T+1}),$$

and show that the conditional distributions

$$G_T(y_1^T; k) = \text{pr}(W_T \leq y_1^T \mid X_{T+1} = k) \quad (k = 1, \dots, K)$$

are linearly independent, and so are

$$H_T(y_1^T; k) = \text{pr}(V_T \leq y_1^T \mid X_{T+1} = k) \quad (k = 1, \dots, K). \quad (2)$$

This is the crucial non-obvious step in our setting, since the state-dependent distribution functions F_1, \dots, F_K may be linearly dependent, and it requires some technical effort. It is essential to make the arguments in Allman et al. [1] work.

In the second step, we follow Allman et al. [1], who rely on Theorem 4a of Kruskal [12] and conclude that for $T \geq K - 1$ the distribution functions $H_T(\cdot; k), F_k, G_T(\cdot; k)$ ($k = 1, \dots, K$) are identified up to joint label swapping. The linear independence from step 1 together with the assumption that the F_k are all distinct will result in a sum of Kruskal ranks not less than $2K + 2$, see the supplement, as required for the argument in this step.

In the third step, we relate the identified distributions $G_{K-1}(\cdot; k)$ and $G_K(\cdot; l)$ ($k, l = 1, \dots, K$) via the transition probability matrix Γ which will thus also be identified. \square

Proof of Theorem 2. To identify the $H_T(\cdot; k)$ ($k = 1, \dots, K$) in (2), we consider the time reversal

$$\{(X_{T+1}, Y_{T+1}), \dots, (X_1, Y_1)\},$$

which is a segment of a hidden Markov model with inhomogeneous underlying Markov chain and state-dependent distributions F_1, \dots, F_K , the Markov chain starting in $\lambda \Gamma^T$. For technical reasons, we first require that Γ and λ have only positive entries, and then relax these assumptions by using higher transitions of order $t_0 = K^2 - 2K + 2$ which results in a transition probability matrix Γ^{t_0} which has strictly positive entries [8], as well as by starting at time t_0 , which results in a starting distribution $\lambda \Gamma^{t_0}$ with positive entries. \square

Proof of Theorem 5 . The existence of the limit as well as its independence from the starting distributions may be deduced from Kingman's subadditive ergodic theorem, as shown in Leroux [14]. To show definiteness, from the construction in Leroux [14], letting

$$\Delta^{K-1} = \{(s_1, \dots, s_K) \in [0, 1]^K : s_1 + \dots + s_K = 1\}$$

denote the $(K - 1)$ -dimensional unit simplex, one obtains a probability measure Q on $\Delta^{K-1} \times \Delta^{K-1}$ such that for $T \geq 2$,

$$\begin{aligned} & T K \{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} \\ &= \int \int g_T(y_1^T; u, \Gamma_0, f_{1,0}, \dots, f_{K,0}) \log \left\{ \frac{g_T(y_1^T; u, \Gamma_0, f_{1,0}, \dots, f_{K,0})}{g_T(y_1^T; v, \Gamma, f_1, \dots, f_K)} \right\} d\nu^{\otimes T}(y_1^T) dQ(u, v). \end{aligned} \quad (3)$$

The inner integral corresponds to the ordinary Kullback–Leibler divergence of the distribution of the segments (Y_1, \dots, Y_T) from two hidden Markov models with parameters $u, \Gamma_0, f_{1,0}, \dots, f_{K,0}$ and $v, \Gamma, f_1, \dots, f_K$, u and v denoting the starting distributions. Non-negativity is then obvious. To show definiteness, choose $T = (2K + 1)(K^2 - 2K + 2) + 1$. From Theorem 2, which implies identification with arbitrary starting distributions, it follows that for distinct parameters $\Gamma_0, f_{1,0}, \dots, f_{K,0}$ and Γ, f_1, \dots, f_K , the inner integral is strictly positive for any values of u and v , and hence so is (3). \square

Proof of Theorem 6. This follows using arguments from convex analysis similar to those in Lindsay [16]. \square

Proof of Theorem 7. To prove the theorem we may follow the arguments in Leroux [14] for the parametric case to obtain the consistency of $\hat{\Gamma}_n$ as well as $d_w(\hat{\mu}_{k,n}, \tilde{\Theta}_{k,0}) \rightarrow 0$ in probability, where

$$\tilde{\Theta}_{k,0} = \{\mu \in \tilde{\Theta} : f_\mu = f_{\mu_{k,0}}\} \quad (k = 1, \dots, K).$$

The main additional issue is to conclude that $f_{\hat{\mu}_{k,n}}(y) \rightarrow f_{k,0}(y)$ if $\tilde{\Theta}_{k,0}$ contains more than a single mixing distribution. Here for fixed $y \in \mathcal{S}$ we use approximation of $\vartheta \mapsto f_\vartheta(y)$ by Lipschitz-continuous functions and the bounded Lipschitz metric on $\tilde{\Theta}$. \square

References

- [1] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, **37** 3099–3132.
- [2] BICKEL, J. P., RITOV, Y. and RYDÉN, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, **26** 1614–1635.
- [3] DANNEMANN, J., HOLZMANN, H. and LEISTER, A. (2014). Semiparametric hidden Markov models: Identifiability and estimation. *WIREs: Computational Statistics*, **6** 418–425.
- [4] DRTON, M. and WILLIAMS, B. (2011). Quantifying the failure of bootstrap likelihood ratio tests. *Biometrika*, **98** 919–934.
- [5] GASSIAT, E., CLEYNEN, A. and ROBIN, S. (2015). Finite state space non parametric hidden Markov models are in general identifiable. *Statistics and Computing* 1–11.

- [6] GASSIAT, E. and ROUSSEAU, J. (2015). Non parametric finite translation hidden Markov models and extensions. *Bernoulli*, **to appear**.
- [7] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, **29** 1233–1263.
- [8] HOLLADAY, J. C. and VARGA, R. S. (1958). On powers of non-negative matrices. *Proceedings of American Mathematical Society*, **9** 631–634.
- [9] HOLZMANN, H. and SCHWAIGER, F. (2014). Hidden Markov models with state-dependent mixtures: minimal representation, model testing and applications to clustering. *Statistics and Computing* 1–16. URL <http://dx.doi.org/10.1007/s11222-014-9481-1>.
- [10] KEMÉNY, J. and SNELL, J. (1960). *Finite Markov chains*. Van Nostrand, New York.
- [11] KIM, A. K. H. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli*, **20** 1802–1818.
- [12] KRUSKAL, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications* 95–138.
- [13] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73** pp. 805–811.
- [14] LEROUX, B. G. (1990). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, **40** 127–143.
- [15] LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20** 1350–1360.
- [16] LINDSAY, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, **11** 86–94.
- [17] ROUEFF, F. C. and RYDÉN, T. (2005). Nonparametric estimation of mixing densities for discrete distributions. *The Annals of Statistics*, **33** 2066–2108.
- [18] TEICHER, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, **32** 244–248.
- [19] VAN DER VAART, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
- [20] VERNET, E. (2015). Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, **9** 717–752.
- [21] VOLANT, S., BÉRARD, C., MARTIN-MAGNIETTE, M.-L. and ROBIN, S. (2014). Hidden Markov models with mixtures as emission distributions. *Statistics and Computing*, **24** 493–504.

Supplementary material for Nonparametric identification and maximum likelihood estimation for hidden Markov models

Grigory Alexandrovich, Hajo Holzmann¹ and Anna Leister

Fakultät für Mathematik und Informatik, Philipps-Universität Marburg, Germany

1 Proofs of the identification results

1.1 Proofs of the main results

For convenience, we recall the notation, the assumptions and the statements of the theorems. A discrete-time hidden Markov model consists of an observed process $(Y_t)_{t \in \mathbb{N}}$ and a latent, unobserved process $(X_t)_{t \in \mathbb{N}}$, such that first, the Y_t are independent given the X_t , second, the conditional distribution of Y_s given the X_t depends on X_s only and third, X_t is a finite-state Markov chain. We assume that X_t is time-homogeneous. The cardinality K of the state space of X_t is called the number of states. The conditional distributions of Y_s given $X_s = k$ ($k = 1, \dots, K$), are called the state-dependent distributions, and we assume that they are independent of s . The entries of the transition probability matrix are denoted by $\Gamma = (\alpha_{j,k})_{j,k=1,\dots,K}$. Further, assume that the Y_t take values in any subset of Euclidean space $\mathcal{S} \subset \mathbb{R}^q$, and denote the distribution functions of the state-dependent distributions by F_k ($k = 1, \dots, K$).

A1. The transition probability matrix $\Gamma = (\alpha_{j,k})_{j,k=1,\dots,K}$ of (X_t) has full rank and is ergodic.

A2. The state-dependent distributions F_k ($k = 1, \dots, K$) are all distinct.

A3. The Markov chain (X_t) is stationary with starting distribution π , the stationary distribution of Γ .

Theorem 1. *For given K , let Γ, F_1, \dots, F_K and $\tilde{\Gamma}, \tilde{F}_1, \dots, \tilde{F}_K$ be two sets of parameters for a K -state hidden Markov model, such that the joint distributions of (Y_1, \dots, Y_{2K+1}) under both sets of parameters are equal. Further, suppose that Γ and F_1, \dots, F_K satisfy Assumptions A1–A3. Then both sets of parameters coincide up to label swapping.*

In order to keep the arguments as transparent as possible, we first prove the following result.

¹Address for correspondence: Prof. Dr. Hajo Holzmann, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerweinstr. D-35032 Marburg, Germany email: holzmann@mathematik.uni-marburg.de, Fon: + 49 6421 2825454

Proposition 2. *Suppose that for a known number of states K , Assumptions A1–A3 are satisfied. Then the parameters Γ and F_1, \dots, F_K are identified from the joint distribution of (Y_1, \dots, Y_{2K+1}) up to label swapping.*

This proposition states that for given K , the parameters Γ and F_1, \dots, F_K are identified within the class of parameters satisfying Assumptions A1–A3. However, from the proofs and exploiting the full strength of Theorem 4a in Kruskal [4], we also obtain Theorem 1, as shown below.

Before we turn to the proof of Proposition 2, we introduce some notation and recall a result of Kruskal [4]. For a vector space V we let $\dim(V)$ denote its dimension. For vectors $v_1, \dots, v_n \in V$ we let $\text{span}\{v_1, \dots, v_n\}$ denote the subspace of V spanned by v_1, \dots, v_n . Further, for numbers $x_1, \dots, x_n \in \mathbb{R}$ we let $\text{diag}(x_1, \dots, x_n)$ denote the n -dimensional diagonal matrix with entries x_1, \dots, x_n . We let $1_K = (1, \dots, 1) \in \mathbb{R}^K$ and $I_K = \text{diag}(1_K)$ denote the K -dimensional unit matrix. For vectors $z_1, z_2 \in \mathbb{R}^T$ we write $z_1 \leq z_2$ if this holds for each coordinate. For a matrix M we let M' denote its transpose. For given matrices $M_i \in \mathbb{R}^{K \times n_i}$ ($n_i \in \mathbb{N}$; $i = 1, 2, 3$) let $[M_1 \ M_2]$ denote the $K \times (n_1 + n_2)$ block matrix and let

$$\langle M_1, M_2, M_3 \rangle(i_1, i_2, i_3) = \sum_{k=1}^K (M_1)_{k,i_1} (M_2)_{k,i_2} (M_3)_{k,i_3} \quad (i_j = 1, \dots, n_j), \quad (1)$$

a three-way array. The Kruskal rank of a matrix $M \in \mathbb{R}^{K \times n}$, denoted $\text{rank}_{Kr} M$, is the maximal j ($j \in \{0, \dots, K\}$), for which each set of j rows in M are linearly independent (as vectors in \mathbb{R}^n). Then Theorem 4a in Kruskal [4] states that if $M_i, N_i \in \mathbb{R}^{K \times n_i}$ ($n_i \in \mathbb{N}$; $i = 1, 2, 3$) are two sets of real matrices such that

$$\langle M_1, M_2, M_3 \rangle = \langle N_1, N_2, N_3 \rangle$$

and

$$\text{rank}_{Kr} M_1 + \text{rank}_{Kr} M_2 + \text{rank}_{Kr} M_3 \geq 2K + 2,$$

then there exists a permutation matrix P and diagonal matrices Λ_i , such that $\Lambda_1 \Lambda_2 \Lambda_3 = I_K$ and $N_i = \Lambda_i P M_i$ ($i = 1, 2, 3$). \diamond

Proof of Proposition 2.

Step 1 (Blocks in the joint distribution and linear independence of conditional distributions). For $T \geq K - 1$ consider

$$V_T = Y_1^T = (Y_1, \dots, Y_T) \text{ and } W_T = Y_{T+2}^{2T+1} = (Y_{T+2}, \dots, Y_{2T+1}).$$

The conditional distribution functions of W_T given $X_{T+1} = k$ ($k = 1, \dots, K$), are given by

$$G_T(y_1^T; k) = \text{pr}(W_T \leq y_1^T \mid X_{T+1} = k) = \sum_{k_1=1}^K \cdots \sum_{k_T=1}^K \alpha_{k,k_1} \prod_{t=2}^T \alpha_{k_{t-1},k_t} \prod_{t=1}^T F_{k_t}(y_t).$$

From Lemma 6 below we have that $G_T(\cdot; k)$ ($k = 1, \dots, K$) are linearly independent functions on \mathcal{S}^T and furthermore, there exist $z_1, \dots, z_K \in \mathcal{S}^T$ for which the $K \times K$ -matrix

$$A_1 = \{G_T(z_t; k)\}_{k,t=1,\dots,K}$$

has full rank K . Here, k is the row index and t the column index. Further, consider the time reversal

$$\tilde{\Gamma} = (\tilde{\alpha}_{j,k})_{j,k=1,\dots,K}, \quad \tilde{\alpha}_{j,k} = \frac{\pi_k \alpha_{k,j}}{\pi_j}.$$

Then for $(y_T, \dots, y_1) \in \mathcal{S}^T$ we have that

$$\begin{aligned} H_T(y_T, \dots, y_1; k) &= \text{pr} \{V_T \leq (y_T, \dots, y_1) \mid X_{T+1} = k\} \\ &= \sum_{k_1=1}^K \cdots \sum_{k_T=1}^K \tilde{\alpha}_{k,k_1} \prod_{t=1}^{T-1} \tilde{\alpha}_{k_t, k_{t+1}} \prod_{t=1}^T F_{k_t}(y_t). \end{aligned}$$

Applying Lemma 6 with $\tilde{\Gamma}$, we conclude that $H_T(\cdot; k)$ ($k = 1, \dots, K$) are linearly independent functions on \mathcal{S}^T and furthermore, there exist $\tilde{z}_1, \dots, \tilde{z}_K \in \mathcal{S}^T$ for which we have the rank K matrix

$$A_2 = \{H_T(\tilde{z}_t; k)\}_{k,t=1,\dots,K}. \quad (2)$$

Step 2 (Kruskal's theorem and identification of conditional distributions). In this step we show that under Assumptions A1 and A2, for $T \geq K - 1$ the distribution functions $H_T(\cdot; k), F_k, G_T(\cdot; k)$ ($k = 1, \dots, K$) are identified up to joint label swapping.

Let $z, \tilde{z} \in \mathcal{S}^T$ and $y \in \mathcal{S}$ be arbitrary points. Set $m = K(K-1)/2$. From Lemma 4 below there exist points $y_j \in \mathcal{S}$ ($j = 1, \dots, m$), such that the $K \times (m+2)$ -matrix

$$M_2 = [\{F_i(y_j)\}_{i=1,\dots,K; j=1,\dots,m}, \{F_i(y)\}_{i=1,\dots,K}, 1_K]$$

has Kruskal rank at least 2. From step 1 the $K \times (K+2)$ -matrices

$$\begin{aligned} M_3 &= [A_1, \{G_T(z; k)\}_{k=1,\dots,K}, 1_K], \quad M_1 = [A_2, \{H_T(\tilde{z}; k)\}_{k=1,\dots,K}, 1_K], \\ \tilde{M}_1 &= \text{diag}(\pi) M_1, \end{aligned}$$

have full rank K , where we use $\pi_k > 0$ ($k = 1, \dots, K$) for \tilde{M}_1 , and therefore

$$\text{rank}_{Kr}(\tilde{M}_1) + \text{rank}_{Kr}(M_2) + \text{rank}_{Kr}(M_3) = 2K + 2. \quad (3)$$

Now we show that the three-dimensional array

$$M = \langle \tilde{M}_1, M_2, M_3 \rangle$$

as defined in (1), is identified from the joint distribution of Y_1^{2T+1} . In the following, we

write $z_{K+1} = z$, $\tilde{z}_{K+1} = \tilde{z}$, $y_{m+1} = y$. We have that

$$\begin{aligned}
M(i, j, r) &= \sum_{k=1}^K \pi_k H_T(\tilde{z}_i; k) F_k(y_j) G_T(z_r; k) \\
&= \sum_{k=1}^K \pi_k \Pr(Y_1^T \leq \tilde{z}_i \mid X_{T+1} = k) \Pr(Y_{T+1} \leq y_j \mid X_{T+1} = k) \Pr(Y_{T+2}^{2T+1} \leq z_r \mid X_{T+1} = k) \\
&= \sum_{k=1}^K \pi_k \Pr(Y_1^T \leq \tilde{z}_i, Y_{T+1} \leq y_j, Y_{T+2}^{2T+1} \leq z_r \mid X_{T+1} = k) \\
&= \Pr(Y_1^T \leq \tilde{z}_i, Y_{T+1} \leq y_j, Y_{T+2}^{2T+1} \leq z_r) \quad (1 \leq i; \quad r \leq K+2; \quad j = 1, \dots, m+2).
\end{aligned} \tag{4}$$

Similarly,

$$\begin{aligned}
M(K+2, j, r) &= \Pr(Y_{T+1} \leq y_j, Y_{T+2}^{2T+1} \leq z_r), \quad M(K+2, m+2, r) = \Pr(Y_{T+2}^{2T+1} \leq z_r), \\
M(i, m+2, r) &= \Pr(Y_1^T \leq \tilde{z}_i, Y_{T+2}^{2T+1} \leq z_r), \quad M(K+2, j, K+2) = \Pr(Y_{T+1} \leq y_j), \\
M(i, j, K+2) &= \Pr(Y_1^T \leq \tilde{z}_i, Y_{T+1} \leq y_j), \quad M(i, m+2, K+2) = \Pr(Y_1^T \leq \tilde{z}_i),
\end{aligned} \tag{5}$$

and $M(K+2, m+2, K+2) = 1$. Evidently, these quantities are identified from the distribution of Y_1^{2T+1} .

Now, using (3) we apply Theorem 4a in Kruskal [4] to show that the matrices \tilde{M}_1, M_2 and M_3 are identified from M up to scaling and permutation, that is there exist a permutation matrix P and diagonal matrices $\Lambda_1, \Lambda_2, \Lambda_3$, such that $\Lambda_1 P \tilde{M}_1, \Lambda_2 P M_2$, and $\Lambda_3 P M_3$ are known and the relationship $\Lambda_1 \Lambda_2 \Lambda_3 = I_K$ holds. Since we know that in the last column of M_2 there are only ones, we obtain the i th diagonal element of the scaling matrix Λ_2 as $(\Lambda_2 P M_2)_{i, K+2}$ ($i = 1, \dots, K$). Similarly we find the matrix Λ_3 . The elements of Λ_1 can then be determined by the relationship $\Lambda_1 \Lambda_2 \Lambda_3 = I_K$. Hence we identified the matrices \tilde{M}_1, M_2 and M_3 up to simultaneous row permutations and therefore the values $H_T(\tilde{z}; k), F_k(y), G_T(z; k)$ at arbitrary points $z, \tilde{z} \in \mathcal{S}^T$ and $y \in \mathcal{S}$. Finally, we show that for distinct values of $z, \tilde{z} \in \mathcal{S}^T$ and $y \in \mathcal{S}$, the matrices P and Λ_3 remain the same, so that there is a joint label swapping. Suppose that for distinct values, we get \tilde{P} and $\tilde{\Lambda}_3$. The matrix $[A_1, 1_K]$, which is the submatrix of both versions of M_3 consisting of the first K columns and the last column, we obtain

$$\Lambda_3 P[A_1, 1_K] = \tilde{\Lambda}_3 \tilde{P}[A_1, 1_K].$$

As above, since the last column of $P[A_1, 1_K] = \Lambda_3^{-1} \tilde{\Lambda}_3 \tilde{P}[A_1, 1_K]$ equals 1_K as well as the diagonal entries of $\Lambda_3^{-1} \tilde{\Lambda}_3$, we get $\Lambda_3^{-1} \tilde{\Lambda}_3 = I_K$ and hence $P[A_1, 1_K] = \tilde{P}[A_1, 1_K]$. Since $[A_1, 1_K]$ has full row rank K , we get $P = \tilde{P}$, as required.

Step 3 (Identification of Γ). It remains to identify the transition probability matrix Γ . We choose $T = K-1$, and after applying the result in step 2, fix a labeling $H_T(\cdot; k), F_k, G_T(\cdot; k)$ ($k = 1, \dots, K$). For $z_1, \dots, z_K \in \mathcal{S}^T$ as in step 1 and $y \in \mathcal{S}$ we consider the $K \times K$ -matrix

$$A = [G_{T+1}\{(y, z_t); k\}]_{k,t=1,\dots,K}.$$

From step 2, $H_{T+1}(\cdot; k), F_k, G_{T+1}(\cdot; k)$ are identified up to joint label swapping and hence so is the matrix A . Since the F_k are all distinct, we may choose the same labeling as the one fixed for $H_T(\cdot; k), F_k, G_T(\cdot; k)$ ($k = 1, \dots, K$). In this case, we have that

$$A = \Gamma \text{diag}\{F_1(y), \dots, F_K(y)\} A_1,$$

where A_1 is defined in step 1. Now choose y large enough so that $F_k(y) \neq 0$ ($k = 1, \dots, K$), so that Γ is identified as

$$\Gamma = A A_1^{-1} \text{diag}[\{F_1(y)\}^{-1}, \dots, \{F_K(y)\}^{-1}],$$

which concludes the proof. □

Proof of Theorem 1. The sets of parameters are denoted by Γ and F_1, \dots, F_K with stationary starting distribution π , and $\tilde{\Gamma}$ and $\tilde{F}_1, \dots, \tilde{F}_K$ with arbitrary starting distribution λ .

Step 1 in the proof of Proposition 2 applies to the parameters Γ and F_1, \dots, F_K . We define the matrices \tilde{M}_1, M_2 and M_3 as in step 2, these satisfy (3). Further, the matrices N_1, N_2 and N_3 as the conditional distribution functions of V_T, Y_{T+1} and W_T given $X_{T+1} = k$ under the parameters $\tilde{\Gamma}, \tilde{F}_1, \dots, \tilde{F}_K$ and λ , evaluated at the same points as for M_1, M_2 and M_3 . Let $\tilde{N}_1 = \text{diag}(\lambda \tilde{\Gamma}^T) N_1$, where we observe that $\lambda \tilde{\Gamma}^T$ is the marginal distribution of X_{T+1} under this parameter set.

Now, (4) and (5) show that under the assumption that both sets of parameters induce the same distribution of Y_1, \dots, Y_{2T+1} ,

$$\langle \tilde{M}_1, M_2, M_3 \rangle = \langle \tilde{N}_1, N_2, N_3 \rangle.$$

For an application of Theorem 4a in Kruskal [4], it suffices that the matrices \tilde{M}_1, M_2 and M_3 satisfy (3), hence there is a $K \times K$ permutation matrix P and diagonal matrices Λ_i ($i = 1, 2, 3$) with $\Lambda_1 \Lambda_2 \Lambda_3 = I_K$, such that

$$M_i = \Lambda_i P N_i \quad (i = 2, 3) \text{ and } \tilde{M}_1 = \Lambda_1 P \tilde{N}_1.$$

Since M_i, N_i ($i = 2, 3$), have only ones in the last column, $\Lambda_2 = \Lambda_3 = I_K$ and hence also $\Lambda_1 = I_K$. It follows that N_3 and \tilde{N}_1 must also have full rank, and that P is uniquely determined, that $\pi = \lambda \tilde{\Gamma}^T$ which are contained in the last column of $\tilde{M}_1 = \tilde{N}_1$, and that the conclusion of step 2 in Proposition 2 holds true. The equality $\Gamma = \tilde{\Gamma}$ follows as in step 3, and since Γ is invertible and $\pi \Gamma^{-1} = \pi$, from $\pi = \lambda \Gamma^T$ we obtain $\pi = \lambda$. □

We let λ denote an arbitrary K -state probability vector.

Theorem 3. *For a known number of states K , let $\lambda, \Gamma, F_1, \dots, F_K$ and $\tilde{\lambda}, \tilde{\Gamma}, \tilde{F}_1, \dots, \tilde{F}_K$ be two sets of parameters for a K -state hidden Markov model, such that the joint distributions of (Y_1, \dots, Y_T) with $T = (2K+1)(K^2-2K+2)+1$, are equal under both sets of parameters. Further, suppose that Γ and F_1, \dots, F_K satisfy Assumptions A1 and A2. Then both sets of parameters coincide up to label swapping.*

Proof.

Step 1 (First assume that λ has only positive entries). We shall show that in this case, from the joint distribution of (Y_1, \dots, Y_{2K+1}) we identify Γ and F_1, \dots, F_K , and the conditional distributions

$$H_T(y_1^T; k) = \text{pr}(Y_1^T \leq y_1^T \mid X_{T+1} = k) \quad (k = 1, \dots, K; T = K-1, K),$$

up to label swapping. To this end we may follow the proofs of Proposition 2 and Theorem 1, and it remains to show that the distribution functions $H_T(\cdot; k)$ ($k = 1, \dots, K$), are linearly independent, where $T = K-1$. The time reversal (X_{T+1}, \dots, X_1) is an inhomogeneous Markov chain, and therefore

$$\{(X_{T+1}, Y_{T+1}), \dots, (X_1, Y_1)\}$$

is a hidden Markov model with inhomogeneous underlying Markov chain and state-dependent distributions F_1, \dots, F_K . In particular,

$$\lambda^{(t)} = \lambda \Gamma^{t-1}, \quad (\tilde{\Gamma}^{(t)})_{i,j} = \frac{\lambda_j^{(t)} \alpha_{j,i}}{\sum_{k=1}^K \lambda_k^{(t)} \alpha_{k,i}} = (\tilde{\alpha}_{i,j}^{(t)})_{i,j=1,\dots,K} \quad (t = 1, \dots, T)$$

and we have that

$$H_T(y_1^T; k) = \sum_{k_1=1}^K \cdots \sum_{k_T=1}^K \tilde{\alpha}_{k,k_T}^T \prod_{t=2}^T \tilde{\alpha}_{k_t,k_{t-1}}^{(t-1)} \prod_{t=1}^T F_{k_t}(y_t).$$

Since all entries in λ are strictly positive, the matrices $\tilde{\Gamma}^{(t)}$ ($t = 1, \dots, T$) all have full rank. The argument in the proof of Lemma 6 applies to show that the $H_T(\cdot; k)$ ($k = 1, \dots, K$), are linearly independent.

Step 2 (Both Γ and λ have only strictly positive entries). We show that in this case all parameters λ , Γ and F_1, \dots, F_K are identified from the joint distribution of (Y_1, \dots, Y_{2K+1}) . It remains to identify λ . We may argue similarly as in step 3 of Proposition 2. For $T = K-1$, we may identify both $H_T(\cdot; k)$ and $H_{T+1}(\cdot; k)$, where we have chosen a fixed equal labelling for both distribution functions. Again, we find $\tilde{z}_1, \dots, \tilde{z}_K \in \mathcal{S}^T$ such that the identified $K \times K$ -matrix A_2 in (2) has full rank K in this situation as well. For $y \in \mathcal{S}$ consider the identified $K \times K$ -matrix

$$[H_{T+1}\{(\tilde{z}_t, y); k\}]_{k,t=1,\dots,K} = \tilde{\Gamma}^{(T+1)} \text{diag}\{F_1(y), \dots, F_K(y)\} A_2,$$

which, for y large enough so that $F_k(y) \neq 0$ ($k = 1, \dots, K$), allows to identify $\tilde{\Gamma}^{(T+1)}$. Therefore, for each j , we identify

$$\frac{\tilde{\alpha}_{j,i}^{(T+1)}}{\alpha_{i,j}} = \frac{\lambda_i^{(T+1)}}{c_j} \quad (i = 1, \dots, K),$$

where c_j is a positive constant. If we fix j , this identifies $\lambda^{(T+1)}$ up to scale. Since $\lambda^{(T+1)}$ is a probability vector, it is identified and since Γ is invertible and identified and $\lambda^{(T+1)} = \lambda \Gamma^T$, λ itself is identified.

Step 3 (Conclusion of the proof). Now we conclude the proof of the theorem. Let $t_0 =$

$K^2 - 2K + 2$. Then from Holladay and Varga [3], Γ^{t_0} has strictly positive entries. Observe that $(Y_{t_0+1}, \dots, Y_{t_0+2K+1})$ is a segment of a hidden Markov model with starting vector $\lambda\Gamma^{t_0}$, which has only positive entries. Using step 1 we therefore identify Γ and F_1, \dots, F_K . Then, using the result in step 2, from

$$(Y_{t_0+1}, Y_{2t_0+1}, \dots, Y_{(2K+1)t_0+1}),$$

which is a segment of a hidden Markov model where the Markov chain starts in $\lambda\Gamma^{t_0}$ and has transition probability matrix Γ^{t_0} , and the state-dependent distributions are F_1, \dots, F_K , we identify $\tilde{\lambda} = \lambda\Gamma^{t_0}$, and therefore also $\lambda = \tilde{\lambda}\Gamma^{-t_0}$.

□

1.2 Technical lemmas for the identification proofs

Lemma 4. *Let G_k ($k = 1, \dots, K$) be distinct distribution functions. Then there exist $y_1, \dots, y_m \in \mathcal{S}$ where $m = K(K-1)/2$ such that the $K \times (m+1)$ matrix*

$$[\{G_i(y_j)\}_{i=1, \dots, K; j=1, \dots, m}, 1_K]$$

has Kruskal rank at least two.

Proof. There exists a $y_{i,j} \in \mathcal{S}$ such that $G_i(y) \neq G_j(y)$ ($i, j \in \{1, \dots, K\}; i < j$). Let y_1, \dots, y_m be points corresponding to the m pairs of indices. □

The next lemma is the key technical result.

Lemma 5. *Let $t \leq K-1$ and $v_1, \dots, v_t \in \mathbb{R}^K$ be linearly independent vectors. Assume that the entries of v_1 are all strictly positive. Let Γ be a $K \times K$ stochastic matrix of full rank and let F_1, \dots, F_K be distinct distribution functions. Then there exist $y \in \mathcal{S}$ and a $j \in \{1, \dots, t\}$ for which, letting*

$$D_y = \text{diag}\{F_1(y), \dots, F_K(y)\},$$

the $K \times (t+1)$ -matrix

$$[\Gamma v_1, \dots, \Gamma v_t, D_y \Gamma v_j]$$

has full rank $t+1$.

Proof. First, we can construct vectors $o^{(1)}, \dots, o^{(K-t)} \in \mathbb{R}^K$ orthogonal to $\text{span}\{\Gamma v_1, \dots, \Gamma v_t\}$, which are of the form

$$o^{(i)} = [o_1^{(i)}, \dots, o_t^{(i)}, 0, \dots, 0, -1, 0, \dots, 0] \quad (i = 1, \dots, K-t),$$

where the -1 is at the $(t+i)$ th place, after possibly relabeling the coordinates of \mathbb{R}^K . Indeed, observe that the $K \times t$ matrix $\Gamma[v_1, \dots, v_t]$ has rank t , so that there are t linearly

independent rows. Denote by M the $t \times t$ matrix formed from these rows, and by N the $(K - t) \times t$ matrix consisting of the remaining rows, and assume after relabeling that

$$\Gamma[v_1, \dots, v_t] = \begin{bmatrix} M \\ N \end{bmatrix}.$$

For $e_i \in \mathbb{R}^{K-t}$ the i th unit vector, we may set $o^{(i)} = [e'_i N M^{-1}, -e'_i]'$. Now, if there exist $y \in \mathcal{S}$ for which $(D_y \Gamma v_j)' o^{(i)} \neq 0$ for some $i \in \{1, \dots, K - t\}$ and $j \in \{1, \dots, t\}$, then $D_y \Gamma v_j$ cannot be contained in the t -dimensional subspace $\text{span}\{\Gamma v_1, \dots, \Gamma v_t\}$ of \mathbb{R}^K , and the assertion of the lemma follows.

Thus assume that

$$(D_y \Gamma v_j)' o^{(i)} = 0 \quad (y \in \mathcal{S}; i \in \{1, \dots, K - t\}; j \in \{1, \dots, t\}), \quad (6)$$

this will lead to a contradiction. Let $\gamma_1, \dots, \gamma_K$ denote the row vectors of Γ . Set

$$S_i = \text{span}\{o_1^{(i)} F_1(y) \gamma_1 + \dots + o_t^{(i)} F_t(y) \gamma_t - F_{t+i}(y) \gamma_{t+i} \mid y \in \mathcal{S}\} \quad (i = 1, \dots, K - t).$$

Then (6) implies that

$$\text{span}\{S_1, \dots, S_{K-t}\} \subseteq \text{span}\{v_1, \dots, v_t\}^\perp. \quad (7)$$

We first argue that if (7) holds,

$$\dim S_i \geq 2 \quad (i = 1, \dots, K - t). \quad (8)$$

To this end we assert that among the first t elements of $o^{(i)}$ there is at least one non-zero entry. Indeed, suppose that all t entries were equal zero, then by the construction of $o^{(i)}$, definition of S_i and (7), we get that

$$F_{t+i}(y) \gamma_{t+i} v_1 = 0, \quad y \in \mathcal{S},$$

a contradiction since $\gamma_{t+i} v_1 > 0$ and since we assume that v_1 has strictly positive entries.

Thus, assume that $j \in \{1, \dots, t\}$ is such that $o_j^{(i)} \neq 0$. Since F_j and F_{t+i} are distinct distribution functions, there exist $y_1^{(i)}, y_2^{(i)}$ such that the vectors

$$[F_j(y_1^{(i)}), F_{t+i}(y_1^{(i)})]', \quad [F_j(y_2^{(i)}), F_{t+i}(y_2^{(i)})]'$$

are linearly independent, and hence so are the vectors

$$[o_1^{(i)} F_1(y_l^{(i)}), \dots, o_t^{(i)} F_t(y_l^{(i)}), -F_{t+i}(y_l^{(i)})]' \quad (l = 1, 2),$$

of coefficients of the linearly independent vectors $\gamma_1, \dots, \gamma_t, \gamma_{t+i}$, which shows (8). To conclude the proof, we observe that due to the linear independence of $\gamma_1, \dots, \gamma_K$ and the definition of the S_i , we have that

$$S_i \not\subseteq \text{span}\left\{ \bigcup_{j=1, j \neq i}^{K-t} S_j \right\} \quad (i = 1, \dots, K - t).$$

Together with (8) we obtain that

$$\dim(\text{span}\{S_1, \dots, S_{K-t}\}) \geq K - t + 1,$$

a contradiction to (7). This concludes the proof of the lemma. \square

Lemma 6. *Under Assumptions A1 and A2, for $T \geq K - 1$ the conditional distributions of W_T given $X_{T+1} = k$ ($k = 1, \dots, K$), that is the functions $G_T(\cdot; k)$, are linearly independent, and furthermore, there exist $z_1, \dots, z_K \in \mathcal{S}^T$ such that the matrix*

$$A_1 = \{G_T(z_t; k)\}_{k,t=1,\dots,K}$$

has full rank K .

Proof. We show the claim for $T = K - 1$. Since marginal distributions of linearly dependent distributions remain linearly dependent, linear independence then follows for any $T \geq K - 1$, and the existence of corresponding points $z_1, \dots, z_K \in \mathcal{S}^T$ follows from Lemma 17 in Allman et al. [1]. Consider

$$\begin{aligned} \tilde{G}_t(y_1^t; k) &= F_k(y_1) \sum_{k_2=1}^K \cdots \sum_{k_t=1}^K \alpha_{k,k_2} \prod_{s=2}^{t-1} \alpha_{k_s, k_{s+1}} \prod_{s=2}^t F_{k_s}(y_s) \\ &= F_k(y_1) \gamma_k D_{y_2} \Gamma \cdots \Gamma D_{y_t} 1_K \quad (k = 1, \dots, K, \ t \in \{1, \dots, K - 1\}), \end{aligned}$$

where as above, $D_y = \text{diag}\{F_1(y), \dots, F_K(y)\}$ and γ_k are the row vectors of Γ . Since

$$\tilde{A}_1 = \{\tilde{G}_{K-1}(z_t; k)\}_{k,t=1,\dots,K} = \Gamma A_1,$$

it is enough to show the claim of the lemma for \tilde{A}_1 . We proceed by induction and show that there exist

$$z_1^{(t)}, \dots, z_{t+1}^{(t)} \in \mathcal{S}^t \quad (t = 1, \dots, K - 1),$$

for which the vectors

$$v_j^{(t)} = [\tilde{G}_t(z_j^{(t)}; 1), \dots, \tilde{G}_t(z_j^{(t)}; K)] \quad (j = 1, \dots, t + 1),$$

are linearly independent, and $v_1^{(t)}$ has only strictly positive entries. The case $t = K - 1$ will then establish Lemma 6.

Indeed, since the distribution functions are distinct, for $t = 1$ we find $y_1^{(1)}, y_2^{(1)} \in \mathcal{S}$ for which

$$v_j^{(1)} = [F_1(y_j^{(1)}), \dots, F_K(y_j^{(1)})]' \quad (j = 1, 2),$$

are linearly independent, and for which $v_1^{(1)}$ has only positive entries.

Now, suppose that the claim is valid for $t \leq K - 1$. We apply Lemma 5 and find a $y_0 \in \mathcal{S}$ and a $j \in \{1, \dots, t + 1\}$ for which the $K \times (t + 2)$ matrix

$$M = [\Gamma v_1^{(t)}, \dots, \Gamma v_{t+1}^{(t)}, D_{y_0} \Gamma v_j^{(t)}]$$

has full rank $t + 2$, which means that it has a $(t + 2) \times (t + 2)$ submatrix of non-zero determinant. Since $D_y \rightarrow I_K$, as $y \rightarrow \infty$,

$$[D_y \Gamma v_1^{(t)}, \dots, D_y \Gamma v_{t+1}^{(t)}, D_{y_0} \Gamma v_j^{(t)}] \rightarrow M, \quad y \rightarrow \infty,$$

and the corresponding submatrix will also be of non-zero determinant for an appropriate $y \in \mathcal{S}$. The claim for $t + 1$ now follows by setting

$$z_s^{(t+1)} = [y, (z_s^{(t)})] \quad (s = 1, \dots, t + 1), \quad z_{t+2}^{(t+1)} = [y_0, (z_j^{(t)})],$$

which concludes the proof. \square

2 Proofs for nonparametric maximum likelihood estimation

2.1 Proofs of the main results

Let \mathcal{D} be a class of densities on \mathcal{S} with respect to some σ -finite measure ν . Suppose that (Y_t, X_t) is a K -state hidden Markov model with transition probability matrix Γ_0 satisfying Assumptions A1 and A3 and having stationary distribution π_0 , and that the state-dependent distributions $F_{1,0}, \dots, F_{K,0}$ are all distinct and have densities $f_{1,0}, \dots, f_{K,0}$ from the class \mathcal{D} .

For parameters $\lambda, \Gamma, f_1, \dots, f_K$, $n \in \mathbb{N}$ and $y_1^n \in \mathcal{S}^n$ consider

$$g_n(y_1^n; \lambda, \Gamma, f_1, \dots, f_K) = \sum_{x_1=1}^K \cdots \sum_{x_n=1}^K \lambda_{x_1} f_{x_1}(y_1) \prod_{i=2}^n \alpha_{x_{i-1}, x_i} f_{x_i}(y_i),$$

the joint density of n observations under these parameters, and denote the log-likelihood function of Y_1, \dots, Y_n by

$$L_n(\lambda, \Gamma, f_1, \dots, f_K) = \log g_n(Y_1^n; \lambda, \Gamma, f_1, \dots, f_K).$$

A4. The true densities $f_{j,0} \in \mathcal{D}$ satisfy $E\{|\log f_{j,0}(Y_1)|\} < \infty$, $(j = 1, \dots, K)$;

A5. The model satisfies $E\{\log f(Y_1)\}^+ < \infty$, $(f \in \mathcal{D})$.

Theorem 7. Suppose that (Y_t, X_t) is a K -state hidden Markov model with transition probability matrix Γ_0 satisfying Assumptions A1 and A3, and that the state-dependent distributions $F_{1,0}, \dots, F_{K,0}$ are all distinct and have densities $f_{1,0}, \dots, f_{K,0}$ from the class \mathcal{D} , and satisfy Assumption A4. Let λ, λ_0 be K -state probability vectors with strictly positive entries. Under Assumption A5, given $f_1, \dots, f_K \in \mathcal{D}$ we have almost surely as $n \rightarrow \infty$ that

$$\begin{aligned} n^{-1} \{L_n(\lambda, \Gamma, f_1, \dots, f_K) - L_n(\lambda_0, \Gamma_0, f_{1,0}, \dots, f_{K,0})\} \\ \rightarrow -K\{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} \in (-\infty, 0], \end{aligned}$$

and $K\{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} = 0$ if and only if the two sets of parameters are equal up to label swapping.

Proof. The existence of the limit and its independence from the starting distributions may be deduced from Kingman's subadditive ergodic theorem as shown in Leroux [5]. To show definiteness, we briefly recall a construction from Leroux [5]. For a sequence (y_n) in \mathcal{S} , define sequences

$$u^{(n)}, v^{(n)} \in \Delta^{K-1} = \{(s_1, \dots, s_K)' \in [0, 1]^K : s_1 + \dots + s_K = 1\},$$

by

$$\begin{aligned} u_k^{(1)} &= \pi_{0k}, & u_k^{(n+1)} &= \frac{\sum_{j=1}^K u_j^n f_{0j}(y_n) \alpha_{0,jk}}{\sum_{j=1}^K u_j^n f_{0j}(y_n)} \quad (k = 1, \dots, K; \ n = 1, 2, \dots) \\ v_k^{(1)} &= \pi_{0k}, & v_k^{(n+1)} &= \frac{\sum_{j=1}^K v_j^n f_j(y_n) \alpha_{j,k}}{\sum_{j=1}^K v_j^n f_j(y_n)} \quad (k = 1, \dots, K; \ n = 1, 2, \dots), \end{aligned}$$

where π_0 is the stationary distribution of Γ_0 , and we set $0/0 = 0$. Let $\Omega = \{(y_n, u^{(n)}, v^{(n)})_{n \in \mathbb{N}}\}$. Leroux [5] shows that there is a probability measure on Ω , such that if $Q(u, v)$ denotes the distribution of $(u^{(1)}, v^{(1)})$ under this measure, for any $T \in \mathbb{N}$ we have that

$$\begin{aligned} T K \{(\Gamma_0, f_{1,0}, \dots, f_{K,0}), (\Gamma, f_1, \dots, f_K)\} &= \int_{\Delta^{K-1} \times \Delta^{K-1}} \int_{\mathcal{S}^T} g_T(y_1^T; u, \Gamma_0, f_{1,0}, \dots, f_{K,0}) \\ &\quad \cdot \log \left\{ \frac{g_T(y_1^T; u, \Gamma_0, f_{1,0}, \dots, f_{K,0})}{g_T(y_1^T; v, \Gamma, f_1, \dots, f_K)} \right\} d\nu^{\otimes T}(y_1^T) dQ(u, v). \end{aligned}$$

Since the inner integral is an ordinary Kullback–Leibler divergence of two densities, non-negativity is obvious. To show definiteness, choose $T = (2K + 1)(K^2 - 2K + 2) + 1$. Suppose that the two sets of parameters $\Gamma_0, f_{1,0}, \dots, f_{K,0}$ and Γ, f_1, \dots, f_K are not equal up to label swapping. Then from Theorem 3, for any $u, v \in \Delta^{K-1}$, this Kullback–Leibler divergence

$$\int_{\mathcal{S}^T} g_T(y_1^T; u, \Gamma_0, f_{1,0}, \dots, f_{K,0}) \log \left\{ \frac{g_T(y_1^T; u, \Gamma_0, f_{1,0}, \dots, f_{K,0})}{g_T(y_1^T; v, \Gamma, f_1, \dots, f_K)} \right\} d\nu^{\otimes T}(y_1^T) > 0,$$

which implies definiteness. \square

Suppose that $(f_\vartheta)_{\vartheta \in \Theta}$ is a parametric family of densities on \mathcal{S} with respect to some σ -finite measure, where $\Theta \subset \mathbb{R}^d$ is compact. Let $\tilde{\Theta}$ be the set of Borel probability measures on Θ , endowed with the weak topology it is also a compact set. Assume that the map $(y, \vartheta) \mapsto f_\vartheta(y)$ is continuous on $\mathcal{S} \times \Theta$. Given $\mu \in \tilde{\Theta}$, we let

$$f_\mu(y) = \int_{\Theta} f_\vartheta(y) d\mu(\vartheta)$$

denote the corresponding mixture density. Let $\theta = (\Gamma, \mu_1, \dots, \mu_K) \in G \times \tilde{\Theta} \times \dots \times \tilde{\Theta}$, where G is the compact set of K -state transition probability matrices. Given the sequence of observations Y_1, \dots, Y_n , the log-likelihood function is

$$L_n(\theta) = \log \left\{ \sum_{x_1=1}^K \cdots \sum_{x_n=1}^K \lambda_{x_1} f_{\mu_{x_1}}(Y_1) \prod_{i=2}^n \alpha_{x_{i-1}, x_i} f_{\mu_{x_i}}(Y_i) \right\},$$

where λ is an arbitrary K -state strictly positive probability vector.

Theorem 8. *Let $(f_\vartheta)_{\vartheta \in \Theta}$ be a parametric family of densities with $\Theta \subset \mathbb{R}^d$ compact, and let $\mathcal{D} = \{f_\mu : \mu \in \tilde{\Theta}\}$ be the model for the state-dependent densities, where $\tilde{\Theta}$ is the set of Borel probability measures on Θ . Then, for any $n \geq 1$ there exists a maximum likelihood estimator $\hat{\theta}_n = (\hat{\Gamma}_n, \hat{\mu}_{1,n}, \dots, \hat{\mu}_{K,n})$, for which the state-dependent mixing distributions are of the form*

$$\hat{\mu}_{k,n} = \sum_{j=1}^m a_j \delta_{\vartheta_{j,k}} \quad (k = 1, \dots, K),$$

where $m \in \{1, \dots, Kn + 1\}$, $a_j > 0$, $\sum_{j=1}^m a_j = 1$, $\vartheta_{j,k} \in \Theta$ ($j = 1, \dots, m$) and where δ_ϑ is the point-mass at ϑ .

Proof. Since by assumption, Θ is compact and $f(y)$ is continuous for any $y \in \mathcal{S}$, if $\mu_n \rightarrow \mu$ weakly then $\int_\Theta f_\vartheta(y) \mu_n(d\vartheta) \rightarrow \int_\Theta f_\vartheta(y) \mu(d\vartheta)$. Therefore, the map

$$\begin{aligned} \Psi : \tilde{\Theta} \times \dots \times \tilde{\Theta} &\longrightarrow \mathbb{R}^n \times \dots \times \mathbb{R}^n \\ (\mu_1, \dots, \mu_K) &\longmapsto [\{f_{\mu_1}(y_t)\}_{t=1, \dots, n}, \dots, \{f_{\mu_K}(y_t)\}_{t=1, \dots, n}], \end{aligned}$$

is affine and continuous. Since $\tilde{\Theta}$ is compact as well, the image

$$D = \Psi(\tilde{\Theta} \times \dots \times \tilde{\Theta}) \subset \mathbb{R}^{Kn}$$

is compact and convex. For fixed $\Gamma = (\alpha_{j,k})_{j,k=1, \dots, K}$, we need to consider maximizing

$$\tilde{L}_n(t_1, \dots, t_K) = \log \left(\sum_{x_1=1}^K \dots \sum_{x_n=1}^K \lambda_{x_1} t_{x_1,1} \prod_{j=2}^n \alpha_{x_{j-1}x_j} t_{x_j,j} \right)$$

over D , where $t_k = (t_{k,1}, \dots, t_{k,n})$ and $z = [t_1, \dots, t_K] \in D$. Since D is compact and \tilde{L}_n is continuous, there exists $z^* = (t_1^*, \dots, t_K^*) \in D$, $t_i^* \in \mathbb{R}^n$ ($i = 1, \dots, K$) where \tilde{L}_n is maximal. Since D is also convex, according to Carathéodory's theorem z^* can be expressed by a convex combination of at most $Kn + 1$ extreme points $s_j^* \in D$, so that

$$z^* = \sum_{j=1}^{Kn+1} a_j s_j^*, \tag{9}$$

where the weights $a_j \geq 0$ sum to one. The s_j^* are images of extreme points in $\tilde{\Theta} \times \dots \times \tilde{\Theta}$ under the affine map Ψ [see 6]. Further, points in the Cartesian product $\tilde{\Theta} \times \dots \times \tilde{\Theta}$ are extreme if and only if all coordinates are extreme in $\tilde{\Theta}$, and the extreme points in $\tilde{\Theta}$ are given by the point masses δ_ϑ for a $\vartheta \in \Theta$. Therefore, there exist $\vartheta_{k,j} \in \Theta$ ($k = 1, \dots, K$; $j = 1, \dots, nK + 1$) such that

$$\Psi(\delta_{\vartheta_{1,j}}, \dots, \delta_{\vartheta_{K,j}}) = s_j^*.$$

Let $m \in \{1, \dots, Kn + 1\}$ denote the number of extreme points needed in the convex

combination (9) for which $a_j > 0$. Then, after relabeling we obtain

$$z^* = \sum_{j=1}^m a_j s_j^* = \sum_{j=1}^m a_j \Psi(\delta_{\vartheta_{1,j}}, \dots, \delta_{\vartheta_{K,j}}) = \Psi\left(\sum_{j=1}^m a_j \delta_{\vartheta_{1,j}}, \dots, \sum_{j=1}^m a_j \delta_{\vartheta_{K,j}}\right).$$

Since

$$\sup_{(\Gamma, \mu_1, \dots, \mu_K)} L_n(\theta) = \sup_{\Gamma} \sup_{\mu_1, \dots, \mu_K} L_n(\theta)$$

the theorem follows. \square

Assume that the true state-dependent densities $f_{k,0} = f_{\mu_{k,0}}$ belong to the model and are all distinct, and that $\Gamma_0 \in G$ satisfies Assumption A1.

A6. For every $\mu \in \tilde{\Theta}$ and a small enough neighborhood O_μ of μ we have

$$E\left[\sup_{\tilde{\mu} \in O_\mu} \{\log f_{\tilde{\mu}}(Y_1)\}^+\right] < \infty.$$

Theorem 9. Let $(f_\vartheta)_{\vartheta \in \Theta}$ be a parametric family of densities with $\Theta \subset \mathbb{R}^d$ compact, and let $\mathcal{D} = \{f_\mu : \mu \in \tilde{\Theta}\}$ be the model for the state-dependent densities, where $\tilde{\Theta}$ is the set of Borel probability measures on Θ . Suppose that Assumptions A1, A3, A4 and A6 hold and let $\hat{\theta}_n = (\hat{\Gamma}_n, \hat{\mu}_{1,n}, \dots, \hat{\mu}_{K,n})$ denote a maximum likelihood estimator. Then, after relabeling, we have in probability as $n \rightarrow \infty$ that $\hat{\Gamma}_n \rightarrow \Gamma_0$ and for any $y \in \mathcal{S}$ and $k \in \{1, \dots, K\}$ that

$$f_{\hat{\mu}_{k,n}}(y) \rightarrow f_{k,0}(y).$$

Furthermore, if the mixing distribution μ is identified from the mixture density f_μ , then we have additionally that $d_w(\hat{\mu}_{k,n}, \mu_{k,0}) \rightarrow 0$ in probability, where d_w is a distance which metrizes weak convergence in $\tilde{\Theta}$.

Proof. We let

$$\tilde{\Theta}_{k,0} = \{\mu \in \tilde{\Theta} : f_\mu = f_{\mu_{k,0}}\} \quad (k = 1, \dots, K).$$

Moreover, for the proof we set

$$\Lambda = G \times \tilde{\Theta} \times \dots \times \tilde{\Theta} \quad \text{and} \quad \Lambda_0 = \{\Gamma_0\} \times \tilde{\Theta}_{1,0} \times \dots \times \tilde{\Theta}_{K,0}.$$

We shall metrize weak convergence on $\tilde{\Theta}$ using the bounded Lipschitz metric [7]

$$d_{BL}(\mu_1, \mu_2) = \sup \left\{ \left| \int f d\mu_1 - \int f d\mu_2 \right|; f : \Theta \rightarrow [0, 1], |f(\vartheta_1) - f(\vartheta_2)| \leq d(\vartheta_1, \vartheta_2) \right\}.$$

On G we take any metric equivalent to the Euclidean metric and on Λ , we take the product metric which we denote by d . The proof consists of two steps.

Step 1. Show that in probability as $n \rightarrow \infty$,

$$d(\hat{\theta}_n, \Lambda_0) \rightarrow 0,$$

which in particular implies $\hat{\Gamma}_n \rightarrow \Gamma_0$.

Step 2. Show that from the convergence in probability

$$d_{BL}(\hat{\mu}_{n,k}, \tilde{\Theta}_{k,0}) \rightarrow 0$$

it follows that for any $y \in \mathcal{S}$, in probability

$$f_{\hat{\mu}_{n,k}}(y) \rightarrow f_{k,0}(y).$$

Consider first step 2. Since Θ is compact, the function $\vartheta \mapsto f_{\vartheta}(y) = g(\vartheta)$ is uniformly continuous and bounded. Therefore, given ε , from Lemma 10 below there is a Lipschitz-continuous h such that $|g(\vartheta) - h(\vartheta)| < \varepsilon$. Let $K_1(\varepsilon) = \sup_{\vartheta \in \Theta} |h(\vartheta)|$ and let $K_2(\varepsilon)$ denote the Lipschitz-constant of h , and let $K(\varepsilon) = \max\{K_1(\varepsilon), K_2(\varepsilon)\}$.

Given any $\mu \in \tilde{\Theta}$, there is a $\nu \in \tilde{\Theta}_{k,0}$ for which $d_{BL}(\mu, \nu) \leq d_{BL}(\mu, \tilde{\Theta}_{k,0}) + \varepsilon/K(\varepsilon)$. From the definition of $\tilde{\Theta}_{k,0}$,

$$f_{k,0}(y) = \int g(\vartheta) d\nu(\vartheta).$$

Therefore, we may estimate

$$\begin{aligned} |f_{\mu}(y) - f_{k,0}(y)| &= \left| \int g(\vartheta) d\mu(\vartheta) - \int g(\vartheta) d\nu(\vartheta) \right| \\ &\leq \int |g(\vartheta) - h(\vartheta)| d\mu(\vartheta) + \left| \int h(\vartheta) d\mu(\vartheta) - \int h(\vartheta) d\nu(\vartheta) \right| \\ &\quad + \int |h(\vartheta) - g(\vartheta)| d\nu(\vartheta) \\ &\leq 2\varepsilon + K(\varepsilon) d_{BL}(\mu, \nu) \leq 3\varepsilon + K(\varepsilon) d_{BL}(\mu, \tilde{\Theta}_{k,0}). \end{aligned}$$

Letting $\delta = \varepsilon/K(\varepsilon)$, we therefore obtain

$$\text{pr} \{ |f_{\hat{\mu}_{n,k}}(y) - f_{k,0}(y)| > 4\varepsilon \} \leq \text{pr} \{ d_{BL}(\hat{\mu}_{n,k}, \tilde{\Theta}_{k,0}) > \delta \} \rightarrow 0.$$

For step 1. we may follow the argument in Leroux [5]. For a parameter vector $\theta = (\Gamma, \mu_1, \dots, \mu_K)$ set

$$\begin{aligned} M_{s,t}(\theta) &= \max_{k \in \{1, \dots, K\}} \{ f_{\mu_k}(Y_{s+1}) \sum_{x_2=1}^K \cdots \sum_{x_{t-s}=1}^K \alpha_{k,x_2} f_{\mu_{x_2}}(Y_{s+2}) \\ &\quad \prod_{i=3}^{t-s} \alpha_{x_{i-1}, x_i} f_{\mu_{x_i}}(Y_{s+i}) \} \quad (s, t \in \mathbb{N}_0; s < t). \end{aligned}$$

From Leroux [5, Lemma 3] we have

$$M_{s,t}(\theta) \leq M_{s,u}(\theta) M_{u,t}(\theta) \quad (s < u < t),$$

so that the process $\log M_{s,t}(\theta)$ is subadditive. From Kingmans' subadditive ergodic theorem,

$$n^{-1} \log M_{0,n}(\theta) \rightarrow H(\theta_0, \theta)$$

in L_1 and almost surely, and from the arguments in Leroux [5] and Theorem 7,

$$H(\theta_0, \theta) \leq H(\theta_0, \theta_0), \quad H(\theta_0, \theta) = H(\theta_0, \theta_0) \Leftrightarrow \theta \in \Lambda_0.$$

Similarly, for a $\theta \in \Theta$ and a neighborhood O_θ of θ , the process $\log \sup_{\theta \in O_\theta} M_{s,t}(\theta)$ is subadditive as well, and the limit

$$n^{-1} \log \sup_{\theta \in O_\theta} M_{0,n}(\theta) \rightarrow H(\theta_0, \theta, O_\theta).$$

The argument in Leroux [5] shows that there exists a $\delta > 0$, such that for every $\theta \in \Lambda \setminus \Lambda_0$ there is a neighborhood O_θ for which

$$H(\theta_0, \theta, O_\theta) \leq H(\theta_0, \theta_0) - \delta.$$

Given $\varepsilon > 0$ let $\Lambda_\varepsilon = \{\lambda \in \Lambda, d(\lambda, \Lambda_0) \geq \varepsilon\}$. Since Λ is compact and the distance is continuous, Λ_ε is compact. Therefore, we can find finitely many O_{θ_j} ($j = 1, \dots, m$), which cover Λ_ε . Since $L_n(\theta)$ and $\log M_{0,n}(\theta)$ have the same asymptotics, we obtain

$$n^{-1} \sup_{\theta \in \Lambda_\varepsilon} L_n(\theta) \leq n^{-1} \max_{j=1, \dots, m} \sup_{\theta \in O_{\theta_j}} L_n(\theta) \rightarrow \max_{j=1, \dots, m} H(\theta_0, \theta_j, O_{\theta_j}) \leq H(\theta_0, \theta_0) - \delta.$$

Since $n^{-1} L_n(\theta_0) \rightarrow H(\theta_0, \theta_0)$, and

$$\{\hat{\theta}_n \in \Lambda_\varepsilon\} \subset \left\{ n^{-1} \sup_{\theta \in \Lambda_\varepsilon} L_n(\theta) \geq n^{-1} L_n(\theta_0) \right\},$$

we obtain $\text{pr}(\hat{\theta}_n \in \Lambda_\varepsilon) \rightarrow 0$. □

2.2 Proof of an additional lemma

In the proof of Theorem 9, we used the following well-known lemma, for which, using argument in Garrido and Jaramillo [2], we provide a proof for convenience of the reader.

Lemma 10. *Let (Θ, d) be an arbitrary metric space, not necessarily compact. Every bounded and uniformly continuous function g on Θ can be uniformly approximated by Lipschitz-continuous functions.*

Proof. If g is bounded and uniformly continuous, then so are its positive and negative part. Therefore, we may assume that the given function $g \geq 0$. Choose $M > 0$ such that $|g(\vartheta)| < M$ for all $\vartheta \in \Theta$, and given $\varepsilon > 0$ let $N \in \mathbb{N}$ such that $(N+1)\varepsilon \geq M$. Define the sets

$$C_n = \{\vartheta \in \Theta : (n-1)\varepsilon < g(\vartheta) < (n+1)\varepsilon\} \quad (n = 0, 1, \dots, N),$$

which cover Θ . Evidently, $C_n \cap C_m = \emptyset$ for $|n-m| > 1$. Using the uniform continuity of g , we may choose $\delta > 0$ such that $|g(\eta) - g(\vartheta)| < \varepsilon/2$ whenever $d(\eta, \vartheta) < \delta$. First we show that for every $\vartheta \in \Theta$ there is a $m \in \{0, \dots, N\}$ with

$$B_\delta(\vartheta) = \{\eta \in \Theta : d(\vartheta, \eta) < \delta\} \subseteq C_m. \quad (10)$$

For the proof, first observe that if ϑ is contained in just a single set C_m , we must have $g(\vartheta) = m\varepsilon$, and $B_\delta(\vartheta) \subseteq C_m$ is obvious by the choice of δ and the definition of C_m . Now

suppose that $\vartheta \in C_n \cap C_{n+1}$ for some $n \in \{0, \dots, N-1\}$, so that $n\varepsilon < g(\vartheta) < (n+1)\varepsilon$. If $n\varepsilon < g(\vartheta) \leq (n+1/2)\varepsilon$ we take $m = n$, otherwise we take $m = n+1$, then (10) follows.

Now define the functions

$$g_n(\vartheta) = \inf\{1, d(\vartheta, \Theta \setminus C_n)\} \in [0, 1],$$

where $d(\vartheta, \emptyset) = \infty$. The g_n are supported on C_n and Lipschitz continuous with constant 1, since for $\vartheta_1 \neq \vartheta_2$

$$\frac{|g_n(\vartheta_1) - g_n(\vartheta_2)|}{d(\vartheta_1, \vartheta_2)} \leq \frac{|d(\vartheta_1, \Theta \setminus C_n) - d(\vartheta_2, \Theta \setminus C_n)|}{d(\vartheta_1, \vartheta_2)} \leq \frac{d(\vartheta_1, \vartheta_2)}{d(\vartheta_1, \vartheta_2)} = 1.$$

Define $h(\vartheta) = \sum_{n=0}^N g_n(\vartheta)$. From (10), we have that $h(\vartheta) \geq \delta$ for all $\vartheta \in \Theta$, and since each ϑ can at most be contained in two sets C_n, C_{n+1} , we have $h(\vartheta) \leq 2$. Further, for $\vartheta_1, \vartheta_2 \in \Theta$ we have

$$|h(\vartheta_1) - h(\vartheta_2)| \leq \sum_{n=0}^N |g_n(\vartheta_1) - g_n(\vartheta_2)| \leq (N+1)d(\vartheta_1, \vartheta_2),$$

which proves that h is a Lipschitz function with constant $(N+1)$. Now, set $\tilde{h}(\vartheta) = h(\vartheta)^{-1} \sum_{n=0}^N n g_n(\vartheta)$. We shall show that \tilde{h} is also Lipschitz continuous and that

$$\sup_{\vartheta \in \Theta} |g(\vartheta) - \varepsilon \tilde{h}(\vartheta)| \leq 2\varepsilon. \quad (11)$$

To this end, we compute

$$\begin{aligned} |\tilde{h}(\vartheta_1) - \tilde{h}(\vartheta_2)| &\leq \sum_{n=0}^N \left| \frac{1}{h(\vartheta_1)} n g_n(\vartheta_1) - \frac{1}{h(\vartheta_2)} n g_n(\vartheta_2) \right| \\ &\leq \sum_{n=0}^N \frac{n g_n(\vartheta_1) |h(\vartheta_2) - h(\vartheta_1)|}{h(\vartheta_1) h(\vartheta_2)} + \sum_{n=0}^N \frac{n |g_n(\vartheta_1) - g_n(\vartheta_2)|}{h(\vartheta_2)} \\ &\leq \{(N+1)^3 \delta^{-2} + (N+1)^2 \delta^{-1}\} d(\vartheta_1, \vartheta_2), \end{aligned}$$

by the properties of h and the g_n . As for (11), suppose that $\vartheta \in C_m$. Then

$$\begin{aligned} |\varepsilon \tilde{h}(\vartheta) - g(\vartheta)| &\leq \varepsilon \left| \frac{(m-1)g_{m-1}(\vartheta) + m g_m(\vartheta) + (m+1)g_{m+1}(\vartheta)}{g_{m+1}(\vartheta) + g_m(\vartheta) + g_{m-1}(\vartheta)} - m \right| + |\varepsilon m - g(\vartheta)| \\ &\leq \varepsilon \left| \frac{g_{m-1}(\vartheta) - g_{m+1}(\vartheta)}{g_{m+1}(\vartheta) + g_m(\vartheta) + g_{m-1}(\vartheta)} \right| + \varepsilon \leq 2\varepsilon. \end{aligned}$$

□

3 Additional simulation results for different numbers of observations

For the simulation scenario considered in section 4 · 1 we report additional simulation results for several choices of the sample size n . Tables 1–5 illustrate the consistency of the nonparametric maximum likelihood estimator. Further, one observes that for series shorter than 1000, the nonparametric maximum likelihood estimator is not superior to the misspecified parametric models in terms of the relative errors.

y	−15.45	−13.77	−11.22	−9.05	−7.26	−5.3	−2.86	−0.21	1.56
nonpar	143.96	40.54	12.65	12.54	22.98	8.22	42.53	46.63	52.72
2-comp	148.17	40.93	11.31	12.19	24.31	8.15	43.72	47.76	55.79
Gauss	162.31	40.59	8.88	10.75	22.67	5.99	41.86	49.97	51.73
y	−9.36	−6.36	−2.71	−0.68	0.5	1.67	3.71	7.36	10.36
nonpar	66.14	34.30	65.12	12.18	17.28	21.57	26.74	68.16	86.82
2-comp	67.78	35.72	68.21	11.64	15.90	20.49	26.90	69.46	88.15
Gauss	74.89	32.72	76.71	9.89	16.80	21.63	27.35	77.79	95.70
y	2.27	3.74	6	7.99	9.66	11.61	14.93	20.17	22
nonpar	1069.26	157.78	13.36	20.92	15.64	10.49	14.01	39.37	54.23
2-comp	1090.58	165.61	11.63	21.49	14.58	9.57	14.08	41.38	55.06
Gauss	1280.50	207.43	5.60	25.13	20.07	8.00	5.33	35.27	50.02

Table 1: Relative errors ($\times 100$) of the three estimators compared to the true densities at selected values for y averaged over 10000 series of length 250. ‘Gauss’ stands for Gaussian state-dependent distributions, ‘2-comp’ for two component Gaussian mixtures and ‘nonpar’ for nonparametric Gaussian mixtures.

y	-15.45	-13.77	-11.22	-9.05	-7.26	-5.3	-2.86	-0.21	1.56
nonpar	125.44	33.44	9.30	12.27	23.50	6.31	43.16	46.72	45.58
2-comp	130.18	33.87	8.43	11.98	24.81	6.28	44.36	48.08	49.25
Gauss	146.53	34.89	7.23	10.50	23.66	5.24	42.57	51.15	41.82
y	-9.36	-6.36	-2.71	-0.68	0.5	1.67	3.71	7.36	10.36
nonpar	64.66	27.17	64.75	10.53	14.94	20.04	25.38	64.05	78.08
2-comp	67.96	28.15	68.57	10.67	14.55	19.63	25.58	65.51	78.84
Gauss	76.33	23.78	75.83	9.58	15.77	20.73	26.07	80.07	97.33
y	2.27	3.74	6	7.99	9.66	11.61	14.93	20.17	22
nonpar	1075.62	161.61	11.29	20.00	14.12	7.96	9.78	35.44	50.19
2-comp	1093.36	170.46	9.72	21.91	14.38	7.26	9.90	38.64	51.30
Gauss	1255.34	204.92	5.11	24.59	19.35	7.24	4.16	34.04	50.41

Table 2: Relative errors ($\times 100$) of the three estimators compared to the true densities at selected values for y averaged over 10000 series of length 500. ‘Gauss’ stands for Gaussian state-dependent distributions, ‘2-comp’ for two component Gaussian mixtures and ‘nonpar’ for nonparametric Gaussian mixtures.

y	-15.45	-13.77	-11.22	-9.05	-7.26	-5.3	-2.86	-0.21	1.56
nonpar	116.05	30.24	7.85	12.60	23.79	5.55	43.60	46.83	44.15
2-comp	122.20	30.72	7.08	12.25	25.06	5.44	44.81	48.33	48.88
Gauss	140.12	32.65	6.40	10.60	24.13	4.89	42.98	52.30	38.85
y	-9.36	-6.36	-2.71	-0.68	0.5	1.67	3.71	7.36	10.36
nonpar	65.00	23.96	64.76	9.97	13.96	19.50	25.26	61.50	72.15
2-comp	68.85	24.75	68.75	10.66	14.07	19.51	25.35	63.35	72.18
Gauss	78.40	19.08	75.38	9.60	15.31	20.32	25.67	81.27	97.87
y	2.27	3.74	6	7.99	9.66	11.61	14.93	20.17	22
nonpar	1082.44	164.38	10.39	20.06	13.86	6.90	8.18	33.98	48.76
2-comp	1098.02	173.48	8.87	22.28	14.67	6.38	8.05	37.87	50.42
Gauss	1242.02	203.51	4.89	24.31	18.97	6.86	3.51	34.35	51.69

Table 3: Relative errors ($\times 100$) of the three estimators compared to the true densities at selected values for y averaged over 10000 series of length 750. ‘Gauss’ stands for Gaussian state-dependent distributions, ‘2-comp’ for two component Gaussian mixtures and ‘nonpar’ for nonparametric Gaussian mixtures.

y	-15.45	-13.77	-11.22	-9.05	-7.26	-5.3	-2.86	-0.21	1.56
nonpar	93.48	22.08	5.20	13.91	23.87	4.15	44.15	43.74	42.45
2-comp	107.55	24.15	4.55	13.14	25.37	3.70	45.70	45.71	52.13
Gauss	131.48	29.27	4.63	10.96	24.73	4.26	43.53	54.12	36.54
y	-9.36	-6.36	-2.71	-0.68	0.5	1.67	3.71	7.36	10.36
nonpar	64.26	18.72	65.32	9.64	12.73	19.32	25.16	55.36	57.27
2-comp	68.83	18.89	68.83	10.82	13.69	19.64	25.35	57.85	57.48
Gauss	81.72	11.59	74.11	9.84	14.53	19.51	25.26	82.88	98.48
y	2.27	3.74	6	7.99	9.66	11.61	14.93	20.17	22
nonpar	1105.51	171.73	9.21	20.85	13.87	5.64	5.01	33.39	48.50
2-comp	1110.86	179.99	7.31	22.93	15.77	5.26	4.55	38.11	51.00
Gauss	1226.72	202.15	4.57	23.86	18.44	6.39	2.53	36.33	54.61

Table 4: Relative errors ($\times 100$) of the three estimators compared to the true densities at selected values for y averaged over 10000 series of length 2000. ‘Gauss’ stands for Gaussian state-dependent distributions, ‘2-comp’ for two component Gaussian mixtures and ‘nonpar’ for nonparametric Gaussian mixtures.

y	-15.45	-13.77	-11.22	-9.05	-7.26	-5.3	-2.86	-0.21	1.56
nonpar	71.64	15.52	3.84	15.11	22.65	3.32	44.27	38.79	39.63
2-comp	97.12	20.84	2.74	14.30	25.50	1.97	46.85	41.65	53.99
Gauss	131.16	29.32	3.44	11.02	24.72	3.92	43.63	54.49	36.31
	-9.36	-6.36	-2.71	-0.68	0.5	1.67	3.71	7.36	10.36
nonpar	64.72	12.76	67.33	9.86	13.44	20.00	24.80	51.26	49.53
2-comp	68.63	13.22	70.28	10.72	14.22	20.43	25.70	54.01	52.37
Gauss	82.78	7.25	74.02	10.10	14.28	19.34	25.53	83.61	98.67
	2.27	3.74	6	7.99	9.66	11.61	14.93	20.17	22
nonpar	1111.82	174.24	8.69	21.15	13.98	5.27	3.54	35.19	51.30
2-comp	1120.81	183.07	6.64	22.93	16.05	4.88	2.99	38.98	52.27
Gauss	1222.89	202.18	4.28	23.57	18.16	6.26	1.83	38.10	56.45

Table 5: Relative errors ($\times 100$) of the three estimators compared to the true densities at selected values for y averaged over 10000 series of length 5000. ‘Gauss’ stands for Gaussian state-dependent distributions, ‘2-comp’ for two component Gaussian mixtures and ‘nonpar’ for nonparametric Gaussian mixtures.

4 An additional simulation scenario with linearly dependent state-dependent distributions

We consider an additional simulation scenario in which the state-dependent distributions are linearly dependent and, moreover, they differ not in location but rather in scale. Let $g_{\beta(a,b)}(x) = \{\Gamma(a+b)\}/\{\Gamma(a)\Gamma(b)\}x^{a-1}(1-x)^{b-1}\mathbb{1}_{(0,1)}(x)$ denote the density of the Beta distribution, and let $g_{\beta(a,b)}(x;l,s) = g_{\beta(a,b)}\{(x-l)/s\}/s$ denote the Beta density translated by l and scaled by s . Again we construct a three-state hidden Markov model with transition probability matrix

$$\Gamma = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.4 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}$$

and let the location parameter μ for the state-dependent mixtures of the first and the second state follow a Beta distribution $g_{\beta(2,11)}(\mu; -3, 20)$, while the scale parameter σ is uniformly distributed on the interval $(0.9, 1.5)$ in the first state and uniformly distributed on the interval $(4, 6)$ in the second state. The density in the third state is a linear combination of those of the first and second states: $f_{3,0}(y) = 0.4f_{1,0}(y) + 0.6f_{2,0}(y)$. We only compare the nonparametric and a parametric Gaussian maximum likelihood estimator. In Fig. 1 we plot estimates of the state-dependent densities and the marginal mixture of the hidden Markov model when using the stationary distributions of the estimated transition probability matrices and the estimated state-dependent densities. In the first state where the density is slightly skew, we observe that the nonparametric maximum likelihood estimator performs better than the parametric estimator, whereas in the second state, where due to the large scale parameters the density is close to being symmetric, the estimators yield similar results. In the third state, the advantage of the nonparametric estimator is obvious, especially in tracing the left tail and the peak of the density.

1st state								
y	-4.31	-2.62	-1.25	-0.17	1.07	3.12	6.35	
nonparametric	22.87	12.85	7.88	18.34	15.61	27.22	72.45	
parametric	27.84	10.48	6.69	19.28	20.16	32.96	94.34	
2nd state								
y	-11.94	-6.67	-2.69	0.07	2.87	7.05	13.66	
nonparametric	20.61	9.25	4.77	8.12	6.59	7.00	40.76	
parametric	21.43	4.92	2.78	4.78	5.38	3.98	33.89	
3rd state								
y	-11.01	-5.06	-1.8	-0.08	1.89	5.57	12.21	
nonparametric	22.97	37.08	15.74	15.93	5.40	22.23	41.73	
parametric	31.77	49.92	20.36	21.09	2.20	30.80	49.30	

Table 6: Relative errors ($\times 100$) of the two estimators compared to the true densities at selected values for y averaged over 10000 replications. ‘Gauss’ stands for Gaussian state-dependent distributions, ‘2-comp’ for two component Gaussian mixtures and ‘nonpar’ for nonparametric Gaussian mixtures.

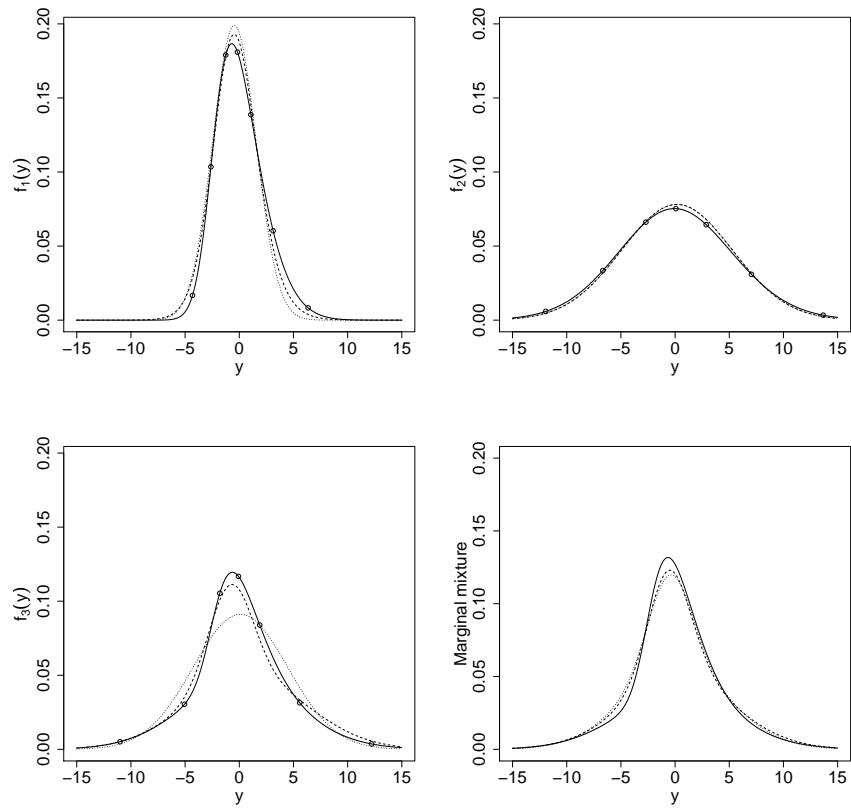


Figure 1: State dependent densities and marginal density of the hidden Markov model. Solid line: true densities, dashed line: nonparametric maximum likelihood estimator, dotted line: Gaussian maximum likelihood estimate.

For the points plotted in Fig. 1 we evaluate the relative errors of the estimators and provide the results averaged over 10000 replications in Table 6. We observe that for the first state, except for two points, the nonparametric estimator yields better results than the parametric estimator. In the second state the parametric estimator yields somewhat better results when estimating the nearly symmetric density. For the third state the nonparametric estimator yields substantially better results than the parametric estimator.

The absolute errors for the estimates of the transition probabilities are reported in Table 7. The nonparametric estimator yields slightly better results in the second and third state.

References

- [1] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, **37** 3099–3132.

	$K^{-1} \sum_{k=1}^K \hat{\alpha}_{j,k} - \alpha_{j,k} $	$K^{-1} \sum_{k=1}^K \tilde{\alpha}_{j,k} - \alpha_{j,k} $
State $j = 1$	11.93	11.71
State $j = 2$	9.65	9.93
State $j = 3$	4.52	5.34

Table 7: Absolute errors of estimated transition probabilities ($\times 100$) averaged over 10000 simulations. Nonparametric estimator ($\hat{\alpha}_{j,k}$) and parametric estimator ($\tilde{\alpha}_{j,k}$) ($j, k = 1, \dots, K$).

- [2] GARRIDO, M. I. and JARAMILLO, J. A. (2008). Lipschitz-type functions on metric spaces. *Journal of Mathematical Analysis and Applications*, **340** 282 – 290.
- [3] HOLLADAY, J. C. and VARGA, R. S. (1958). On powers of non-negative matrices. *Proceedings of American Mathematical Society*, **9** 631–634.
- [4] KRUSKAL, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications* 95–138.
- [5] LEROUX, B. G. (1990). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, **40** 127–143.
- [6] SIMON, B. (2011). *Convexity: An Analytic Viewpoint*. Cambridge University Press.
- [7] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes - With Applications to Statistics*. Springer.